SPECIAL DOCUMENT-ONR 477
July 2003

Fractals Text Mining Using Bibliometrics and Database Tomography

Dr. Ronald N. Kostoff
Dr. Michael F. Schlesinger
ONR

Guido Malpohl
University of Karlsruhe

SPECIAL DOCUMENT-ONR 477
July 2003

Fractals Text Mining Using Bibliometrics and Database Tomography

Dr. Ronald N. Kostoff
Dr. Michael F. Schlesinger
ONR

Guido Malpohl
University of Karlsruhe

CONTENTS

Figure

Tables

1. INTRODUCTION

Science and technology are assuming an increasingly important role in the conduct and structure of domestic and foreign business and government. In the highly competitive civilian and military worlds, there has been a commensurate increase in the need for scientific and technical intelligence to ensure that one's perceived adversaries do not gain an overwhelming advantage in the use of science and technology. While there is no substitute for direct human intelligence gathering, many techniques have become available that can support and complement it. In particular, techniques that identify, select, gather, cull, and interpret large amounts of technological information semi-automatically can expand greatly the capabilities of human beings in performing technical intelligence.

One such technique is Database Tomography (DT) (Kostoff, 1993, 1994, 1995), a system for analyzing large amounts of textual computerized material. It includes algorithms for extracting multi-word phrase frequencies and phrase proximities from the textual databases, coupled with the topical expert human analyst to interpret the results and convert large volumes of disorganized data to ordered information. Phrase frequency analysis (occurrence frequency of multi-word technical phrases) provides the pervasive technical themes of a database, and the phrase proximity (physical closeness of the multi-word technical phrases) analysis provides the relationships among pervasive technical themes, as well as among technical themes and authors/journals/institutions/countries, etc. This paper describes use of the DT process, supplemented by literature bibliometric analyses, to derive technical intelligence from the published literature of Fractals science and technology.

Fractals, as defined by the authors for this study, are geometric structures (e.g., Mandelbrot set, percolation clusters, diffusion-limited aggregates) or dynamical processes (e.g., fractional Brownian motion, avalanches, turbulent intermittency) that possess features on many scales related through a power law relationship. Since one of the key outputs of the present study is a query that can be used by the community to access relevant Fractals documents, a recommended query based on this study is presented in total. This query serves as the operational definition of Fractals, and its development is discussed in detail in the database generation section.

FRACTALS QUERY

FRACTAL* OR SELF-SIMILAR* OR SELF-ORGANIZED CRITICALITY OR MULTIFRACTAL OR ANOMALOUS DIFFUSION OR SCALE INVARIANT OR HAUSDORFF DIMENSION OR DIFFUSION LIMITED AGGREGATION OR FRACTIONAL BROWNIAN MOTION OR MANDELBROT OR LACUNARITY OR CANTOR SET OR NONFRACTAL OR MONOFRACTAL NOT FRACTALKINE*

To execute the study reported in this document, a database of relevant Fractals articles is generated using the iterative search approach of Simulated Nucleation (Kostoff, Eberhart, and Toothman, 1997a; Kostoff et al., 2001). Then, the database is analyzed to produce the following characteristics and key features of the Fractals field: recent prolific Fractals authors; journals that contain numerous Fractals papers; institutions that produce numerous Fractals papers; keywords most frequently specified by the Fractals authors; authors, papers, and journals cited most frequently; pervasive technical themes of Fractals; and relationships among the pervasive themes and sub-themes.

What is the importance of applying DT and bibliometrics to a topical field such as Fractals? The road map, or guide, of this field produced by DT and bibliometrics provides the demographics and a macroscopic view of the total field in the global context of allied fields. This view allows specific starting points to be chosen rationally for more detailed investigations into a specific topic of interest. DT and bibliometrics do not obviate the need for detailed investigation of the literature or interactions with the main performers of a given topical area to make a substantial contribution to the understanding or the advancement of this topical area, but allow these detailed efforts to be executed more efficiently. DT and bibliometrics are quantity-based measures (number of papers published, frequency of technical phrases, etc.), and correlations with intrinsic quality are less direct. The direct quality components of detailed literature investigation and interaction with performers, combined with the DT and bibliometrics analysis, can result in a product highly relevant to the user community.

2. BACKGROUND

2.1 OVERVIEW

Kostoff, Eberhart, and Toothman (1999a) present the information sciences background for the approach used in this paper. This reference shows the unique features of the computer and co-word-based DT process relative to other road-map techniques. It describes the two main road-map categories (expert-based and computer- based), summarizes the different approaches to computer-based road maps (citation and co-occurrence techniques), presents the key features of classical co-word analysis, and shows the evolution of DT from its co-word roots to its present form.

The DT method in its entirety generally requires three distinct steps. The first step is identifying the main themes of the text analyzed. The second step is determinating the quantitative and qualitative relationships among the main themes and their secondary themes. The final step is tracking the evolution of these themes and their relationships through time. The first two steps are summarized in sections 2.1.1 and 2.1.2. Time evolution of themes has not yet been studied.

At this point, a variety of different analyses can be performed. For databases of non-journal technical articles (Kostoff, 1993), the final results have been identification of the pervasive technical themes of the database, the relationship among these themes, and the relationship of supporting sub-thrust areas (high and low frequency) to the high-frequency themes. For the more recent studies in which the databases are journal article abstracts and associated bibliometric information (authors, journals, addresses, etc.), the final results have also included relationships among the technical themes and authors, journals, institutions, etc. (Kostoff et al., 1997b, Kostoff et al., 1998a; Kostoff et al., 2000a; Kostoff et al., 2000b; Kostoff et al., 2002).

These more recent DT/bibliometrics studies were conducted of the technical fields of (1) near-earth space (NES) (Kostoff et al., 1998a), (2) hypersonic and supersonic flow over aerodynamic bodies (HSF) (Kostoff et al., 1999a), (3) chemistry (Kostoff et al., 1997b) as represented by the *Journal of the American Chemical Society* (JSCS), (4) Fullerenes (FUL) (Kostoff et al., 2000a) (5) Aircraft (AIR) (Kostoff et al., 2000b), (6) Hydrodynamic (HYD) flow over surfaces; (7) Electric Power Sources (EPS), (8) Electrochemical Power Sources (ECHEM) (Kostoff et al., 2002), (9) the non-technical field of research impact assessment (RIA) (Kostoff et al., 1997b), and (10) Non-Linear Dynamics (NONLIN) (Kostoff, Shlesinger, and Tshiteya). Table 1 shows the overall parameters of these studies from the Science Citation Index (SCI) database results and the current Fractals study.

Table 1. DT Studies of Topical Fields.

Topical Area	Number of Science Articles	Years Covered
(1) Near-Earth Space (NES)	5480	1993 to Mid-1996
(2) Hypersonics (HSF)	1284	1993 to Mid-1996
(3)Chemistry (JACS)	2150	1994
(4) Fullerenes (FUL)	10,515	1991 to Mid-1998
(5) Aircraft (AIR)	4346	1991 to Mid-1998
(6) Hydrodynamics (HYD)	4608	1991 to Mid-1998
(7) Electric Power Sources (EPS)	20,835	1991 to Begining 2000
(8) Electrochemical Power Sources (ECHEM)	6985	1993 to Mid-2001
(9) Research Assessment (RIA)	2300	1991 to Begining 1995
(10) Non-linear Dynamics (NONLIN)	6118 (2001)	1991, 2001
(11) FRACTALS (FRACT) 4454 (2001–2002); 4211 (1991–1993) 1991–1993; 2001–2002		

2.1.1 First Step

The frequencies of appearance in the total text of all single word phrases (e.g., Matrix), adjacent double word phrases (e.g., Metal Matrix), and adjacent triple word phrases (e.g., Metal Matrix Composites) are computed. The highest frequency significant technical content phrases are selected by topical experts as the pervasive themes of the full database.

2.1.2 Second Step

2.1.2.1 Numerical Boundaries. For each theme phrase, the frequencies of phrases within ±M (nominally, 50) words of the theme phrase are computed for every occurrence of the theme phrase in the full text, and a phrase frequency dictionary is constructed. This dictionary contains the phrases closely related to the theme phrase. Numerical indices are used to quantify the strength of this relationship. Quantitative and qualitative analyses are performed by the topical expert for each dictionary (hereafter called cluster), yielding, among many results, those sub-themes closely related to and supportive of the main cluster theme.

Threshold values are assigned to the numerical indices, and these indices are used to filter out the phrases most closely related to the cluster theme. However, because numbers are limited in their ability to portray the conceptual relationships among themes and sub-themes, the qualitative analyses of the extracted data by the topical experts have been at least as important as the quantitative analyses. The richness and detail of the extracted data in the full-text analysis allow an understanding of the theme inter-relationships not possible with previous text abstraction techniques (using index words, key words, etc.).

2.1.2.2 Semantic Boundaries. The approach is conceptually similar to 2.1.2.1, with the difference that semantic boundaries are used to define the co-occurrence domain rather than numerical boundaries. The only semantic boundaries used for the present studies were paper Abstract boundaries. Software is being developed that will allow paragraphs or sentences to be used as semantic boundaries.

It is an open question as to whether semantic boundaries or numerical boundaries provide more accurate results. The elemental messages of text are contained in concepts or thoughts. Sentences or paragraphs are the vehicles by which the concepts or thoughts are expressed. The goal of text mining is to usually quantify relationships occurring in the concepts or thoughts, not in the fragments of their vehicles of expression. In particular, while intra-sentence relationships will be very strong, they may be overly restrictive for text mining purposes, and many cross-discipline relationships can be lost by adhering to intra-sentence relationships only. Intra-paragraph relationships are more inclusive and reasonable. For journal paper Abstracts of the type found in SCI, many Abstracts constitute a single paragraph.

2.2 UNIQUE STUDY FEATURES

The study reported in this document is in the latter (journal article abstract) category. It differs from the previous published papers in this category (Kostoff et al., 1999a; Kostoff et al., 1998a, 1997b, 2000a, 2000b, 2002) in five respects. First, the topical domain (Fractals) is completely different. Second, a document clustering technique for theme categorization, based on Greedy String Tiling (Wise, 1992) for text similarity, was developed and included to complement the word/ concept clustering approach. Third, bibliometric clustering is presented for two database fields: authors and countries. Fourth, factor matrix filtering was developed and used to select context-dependent words for input to the clustering algorithm, thereby leading to more sharply defined clusters. Finally, the marginal utility algorithm was applied, allowing only the highest payoff terms to be included in the final query, and resulting in an efficient query.

3. DATABASE GENERATION

The key step in the Fractals literature analysis is the generation of the database used for processing. Database generation has three key elements: (1) the overall objectives, (2) the approach selected, and (3) the database used. Each element is described below.

3.1 OVERALL STUDY OBJECTIVES

The main objective was to identify global S&T that had direct and indirect relations to Fractals. A sub-objective was to estimate the overall level of global effort in Fractals S&T, as reflected by the emphases in the published literature.

3.2 DATABASES AND APPROACH

For the present study, the SCI database (including the Science Citation Index and the Social Science Citation Index [SSCI]) was used. The approach used for query development was the DT-based iterative relevance feedback concept (Kostoff et al., 1997a).

3.2.1 SCI/SSCI (SCI, 2002)

The retrieved database used for analysis consists of selected journal records (including the fields of authors, titles, journals, author addresses, author keywords, abstract narratives, and references cited for each paper) obtained by searching the Web version of the SCI for Fractals articles. At the time the final data were extracted for this document (Fall 2002), the version of the SCI used accessed about 5600 journals (mainly in physical, engineering, and life sciences basic research) from the SCI, and over 1700 journals from the SSCI.

The SCI database selected represents a fraction of the available Fractals (mainly research) literature that, in turn, represents a fraction of the Fractals S&T actually performed globally (Kostoff, 2000c). The articles contained within the SCI database do not include the large body of classified literature, or company proprietary technology literature, although the SCI articles could reference this literature. The SCI articles do not include technical reports, books, or patents on Fractals, but could again reference these literatures. The SCI covers a finite slice of time (1991–1993, 2001–2002). The database used represents the bulk of the peer-reviewed high-quality Fractals research literature, and is a recent representative sample of all Fractals research.

To extract the relevant articles from the SCI, the Title, Keyword, and Abstract fields were searched using Keywords relevant to Fractals. The resultant Abstracts were culled to those relevant to Fractals. The search was performed with the aid of two powerful DT tools (multi-word phrase frequency analysis and phrase proximity analysis) using the process of Simulated Nucleation (Kostoff et al., 1997a).

An initial query of Fractals-related terms produced two groups of papers: (1) one group was judged by domain experts as relevant to the subject matter, and (2) the other was judged as non-relevant. Gradations of relevancy or non-relevancy were not considered. An initial database of Titles, Keywords, and Abstracts was created for each of the two groups of papers. Phrase frequency and proximity analyses were performed on this textual database for each group. The high-frequency

single, double, and triple word phrases characteristic of the relevant group, and their Boolean combinations, were then added to the query to expand the papers retrieved. Clustering of phrases into thematic categories was performed to help guide the selection of phrases. Phrases from each thematic category were selected to ensure balanced representation from the complete sample of relevant records. Similar phrases characteristic of the non-relevant group were effectively subtracted from the query to contract the papers retrieved. The process was repeated on the new database of Titles, Keywords, and Abstracts obtained from the search. A few more iterations were performed until the number of records retrieved stabilized (convergence). The final phrase-based query used for the Fractals study was shown in the Introduction.

To generate an efficient final query, a new process termed Marginal Utility was applied. At the start of the final iteration, a modified query Q1 was inserted into the SCI, and records were retrieved. A sample of these records was then categorized into relevant and non-relevant. Each term in Q1 was inserted into the Marginal Utility algorithm, and the marginal number of relevant and non-relevant records in the sample that the query term would retrieve was computed. Only those terms that retrieved a high ratio of relevant to non-relevant records were retained. Since (by design) each query term had been used to retrieve records from the SCI as part of Q1, the marginal ratio of relevant to non-relevant records from the sample would represent the marginal ratio of relevant to non-relevant records from the SCI. The final efficient query Q2, consisting of the highest marginal utility terms, was shown in the Introduction.

In the Marginal Utility algorithm, terms that co-occur strongly in records with previously selected terms are essentially duplicative from the retrieval perspective, and can be eliminated. Thus, the order in which terms are selected becomes important. An automated query term selection algorithm using Marginal Utility is being developed that will examine all ordering combinations to identify the most efficient query.

The authors believe that queries of these magnitudes and complexities are required when necessary to provide a tailored database of relevant records that encompasses the broader aspects of target disciplines. In particular, if it is desired to enhance the transfer of ideas across disparate disciplines, and thereby stimulate the potential for innovation and discovery from complementary literatures (Kostoff, 1999b), then even more complex queries using Simulated Nucleation may be required.

However, even with queries of this magnitude, not all records will be retrieved. As a point of reference, there were 39 articles with Abstracts published in the journal *Fractals* in 2001, of which 31 (~80%) were retrieved for this study. This retrieval was the highest fraction retrieved for any journal examined. For all the journals examined, some records had insufficient verbiage in their text fields, or had very non-standard verbiage relative to the main topical themes. Either of these problems precluded the query's accessing the record(s). To retrieve records with non-standard, very low frequency terminology from all the journals accessed would require queries that contain thousands of terms. **The reader should think about how many fewer Fractals records would have been accessed with the typical search queries containing about a half-dozen terms, and how author and journal citation rates are negatively impacted by the combination of deficient queries and insufficient verbiage in the record text fields.**

4. RESULTS

The results from the publications bibliometric analyses are presented in section 4.1, followed by the results from the citations bibliometrics analysis in section 4.2. Results from the DT analyses are shown in section 4.3. The SCI bibliometric fields incorporated into the database included, for each paper, the author, journal, institution, keywords, and references.

4.1 PUBLICATION STATISTICS ON AUTHORS, JOURNALS, ORGANIZATIONS, COUNTRIES

The first group of metrics presented is counts of papers published by different entities. These metrics can be viewed as output and productivity measures. They are not direct measures of research quality, although some threshold quality level is inferred, since these papers are published in the (typically) high-caliber journals accessed by the SCI.

4.1.1 Author Frequency Results

For 2001–2002, 4464 papers (4380 of which had Abstracts), 9403 different authors, and 12780 author listings were retrieved. The occurrence of each author's name on a paper is defined as an author listing. While the average number of listings per author is about 1.36, the 19 most prolific authors (see Table 2A) have listings more than an order of magnitude greater than the average. The number of papers listed for each author are those in the database of records extracted from the SCI using the query, not the total number of author papers listed in the source SCI database.

Table 2a. Most Prolific Authors (2001–2002)

(Present Institution Listed).

Author	Institution	Country	# of Papers
Stanley H E	Boston University	USA	15
Huikuri H V	University of Oulu	Finland	14
Wu Z Q	University of Science and Technology China	China	13
Zaslavsky G M	New York University	USA	12
Jin Z Z	Wuhan University	China	11
Makikallio T H	University of Oulu	Finland	11
Sidharth B G	BM Birla Science Centre	India	11
Zou X W	Wuhan University	China	11
Havlin S	Bar-Ilan University	Israel	10
Lau K S	Chinese University of Hong Kong	China	10
Mendes R S	University Estadual de Meringa´	Brazil	10
Tan Z J	Wuhan University	China	10
Tsallis C	Centro Brasileiro de Pesquisas Fisicas	Brazil	10
Bershadskii A	ICAR	Israel	9
Fujita H	Hyogo Pref Inst Ind Res	Japan	9
Lapenna V	Consiglio Nazionale delle Ricerche (CNR)	Italy	9
Sun X	University of Science and Technology China	China	9
Veltri P	University of Calabria	Italy	9

Of the 18 most prolific authors listed in Table 2A, six are from China. In fact, six are from the Far East, two are from the East, two are from the Middle East, two are from Western Europe, two are from Northern Europe, two are from North America, and two are from South America. Thirteen are from universities, and five are from research institutes.

To determine the trends in this regional mix of prolific authors, the same query was applied to 1991–1993 only. Table 2B lists the most prolific authors for 1991–1993.

Table 2b. Most Prolific Authors (1991–1993).

Author	Institution	Country	# of Papers
Meakin P	University of Oslo	Norway	24
Stanley H E	Boston University	USA	23
Havlin S	Bar-Ilan University	Israel	20
Vlad M O	KFA Julich GmbH	Germany	19
Nagatani T	Shizuoka University	Japan	18
Balankin A S	Fe Dzerzhinskii Military Academy	Russia	17
Pietronero L	University of Rome La Sapienza	Italy	16
Feder J	University of Oslo	Norway	15
Jossang T	University of Oslo	Norway	14
Salvarezza R C	National Univeristy of La Plata	Argentina	13
Arvia A J	National University of La Plata	Argentina	12
Procaccia I	Weizmann Institue of Science	Israel	12
Sornette D	University of Nice-Sophia Antipolis	France	12
Bras R I	MIT	Usa	11
Giona M	University of Rome La Sapienza	Italy	11
Milosevic S	University of Belgrade	Yugoslavia	11
Mosolov A B	Politecnic of Turin	Italy	11
Sapoval B	Ecole Polytechnique	France	11

The regional mix of authors has some major differences from the 2001 results. Of the 18 most prolific authors listed in Table 2B, one is from the Far East, two are from the Middle East, two are from North America, two are from South America, six are from Western Europe, three are from Northern Europe, and two are from Eastern Europe. Seventeen are from universities, and one is from a research institute.

Only two names were common to both lists, Stanley and Havlin, and they co-author to a reasonable extent. However, some researchers can have an off-year for a number of reasons, so individual comparisons over 2 years, especially two widely separated years, may not be overly important. More important are country comparisons, and maybe institutional comparisons to some extent. These entities integrate over many individuals, and their performance would be more reflective of national policy. In this regard, the aggregate shift of prolific performers from the European countries in 1991–1993 to those of the East/Far East in 2001–2002 stands out.

4.1.2 Journals Containing Most Fractals Papers

For 2001–2002, 1238 different journals were represented, with an average of 3.61 papers per journal. The journals containing the most Fractals papers (see Table 3A) had more than an order of magnitude more papers than the average.

Table 3a. Journals Containing Most Papers (2001–2002).

Journal	# of Papers
Physical Review E	314
Physica A	151
Chaos Solitons & Fractals	100
Physical Review Letters	91
Physical Review B	82
Fractals-Complex Geometry Patterns and Scaling In Nature and Society	60
Astrophysical Journal	55
Physics Letters A	49
Physical Review D	44
Langmuir	38
Journal of Colloid And Interface Science	37
Journal of Physics A-Mathematical and General	36
Europhysics Letters	34
Astronomy & Astrophysics	33
Journal of Fluid Mechanics	31
Journal of Statistical Physics	29
European Physical Journal B	28
Monthly Notices of the Royal Astronomical Society	28
Physics of Plasmas	26

Essentially all of the journals are physics journals, ranging in mission from dedication to fractals (FRACTALS) to sub-branches of physics that include fractal analyses (PHYSICS OF PLASMAS).

To determine the trends in journals containing the most Fractals papers, the results from 1991–1993 are examined. Table 3B contains the top 20 journals.

Table 3b. Journals Containing Most Papers (1991–1993).

Journal	# of Papers
Physica A	213
Physical Review A	174
Physical Review Letters	173
Physical Review B-Condensed Matter	115
Physical Review E	86
Astrophysical Journal	86
Physics Letters A	85
Journal of Physics A-Mathematical and General	77
Journal of Statistical Physics	73
Physica D	57
Europhysics Letters	52
Physics of Fluids A-Fluid Dynamics	50
Physics Letters B	50
Physical Review D	44
Journal of Physics-Condensed Matter	43
Geophysical Research Letters	40
Journal of Chemical Physics	35
Journal of Non-Crystalline Solids	33
Journal of The Physical Society Of Japan	32
Journal of Fluid Mechanics	32

While the most prolific authors could be expected to change over a decade, for a number of reasons, the most prolific journals should be more stable. Comparison of Tables 3A and 3B shows this is true. Of the 20 most prolific journals, 11 are in common.

The journals in the top 20 in 1991–1993 that were not included in the top 20 from 2001–2002 tended to be the more traditional discipline-oriented physics journals (JOURNAL OF PHYSICS-CONDENSED MATTER, GEOPHYSICAL RESEARCH LETTERS, JOURNAL OF CHEMICAL PHYSICS, JOURNAL OF NON-CRYSTALLINE SOLIDS, PHYSICS OF FLUIDS-FLUID DYNAMICS, ETC). The journals in the top 20 in 2001–2002 that were not included in the top 20 from 1991–1993 tended to be the more generic non-discipline oriented physics journals (FRACTALS, CHAOS SOLITONS AND FRACTALS, LANGMUIR, JOURNAL OF COLLOID AND INTERFACE SCIENCE, ETC). Additionally, some of these journals are relatively new, and that may account for their increasing prominence from 1991 to 2001.

4.1.3 Institutions Producing Most Fractals Papers

A similar process was used to develop a frequency count of institutional address appearances. Note that many different organizational components may be included under the single organizational heading (e.g., Harvard University could include the Chemistry Department, Biology Department, Physics Department, etc.). Identifying the higher level institutions is instrumental for these DT studies. Once they have been identified through bibliometric analysis, subsequent measures may be taken (if desired) to identify particular departments within an institution.

Table 4a. Prolific Institutions (2001–2002).

Institution	Country	# Papers
Russian Academy of Science	Russia	135
Chinese Academy of Science	China	65
MIT	USA	54
University of Cambridge	UK	47
University of Paris	France	46
Centre National de la Recherche Scientifique (CNRS)	France	43
Boston University	USA	42
Consiglio Nazionle delle Ricerche (CNR)	Italy	40
University of Science and Technology China	China	38
University of California Los Angeles	USA	37
University of Tokyo	Japan	35
University of California Berkeley	USA	34
Harvard University	USA	31
Kyoto University	Japan	31
Ecole Polytechnique	France	31
Cornell University	USA	29
Polish Academy of Science	Poland	29
Chinese University Hong Kong	China	28
Tsing Hua University	China	28
Penn State University	USA	28

For 2001–2002, of the 20 most prolific institutions, 7 are from the USA, 5 are from Western Europe, 6 are from Asia, and 2 are from Eastern Europe. Fifteen are universities, and the remaining institutions are research institutes.

To determine the trends in institutions containing the most Fractals papers, the results from 1991–1993 were examined. Table 4B contains the top 20 institutions.

Table 4b. Prolific Institutions (1991–1993).

Institution	Country	# of Papers
Russian Academy of Science	Russia	110
Tel Aviv University	Israel	51
IBM Corporation	USA	49
Cornell University	USA	48
NASA	USA	47
KFA Julich Gmbh	Germany	47
MIT	USA	47
University of Chicago	USA	45
University of Cambridge	UK	45
University of Illinois	USA	45
Acad Sinica	Taiwan/China	44
University of Maryland	USA	44
University of Tokyo	Japan	42
University of California San Diego	USA	40
University of Rome La Sapienza	Italy	39
University of California Berkeley	USA	38
Boston University	USA	35
University of Michigan	USA	34
Princeton University	USA	34
Ecole Polytechnique	France	33

Of the 20 most prolific institutions in 1991–1993, 12 are from the USA, 4 are from Western Europe, 1 is from Eastern Europe, 1 is from the Middle East, and 1 is from Taiwan/China. The major shift is substitution of Asian institutions for USA institutions. Sixteen institutions are universities, four are research institutes, and one is industrial research.

4.1.4 Countries Producing Most Fractals Papers

Ninety different countries are listed in the results for 2001–2002. Table 5A summarizes the country bibliometric results. The dominance of a handful of countries is clearly evident.

Table 5a. Prolific Countries (2001–2002).

Country	# of Papers
USA	1223
France	464
Peoples Republic of China	398
Germany	373
Japan	340
Russia	329
England	299
Italy	277
Spain	172
Canada	167
Brazil	156
Poland	137
India	112
Israel	112
Australia	110
Netherlands	84
Greece	71
Taiwan	69
Sweden	68
South Korea	63
Argentina	60
Switzerland	57
Hungary	56
Belgium	51
Finland	49
Ukraine	47
Denmark	43
Scotland	42
Mexico	41
Austria	37
New Zealand	29

There appears to be two dominant groupings. The first group is the USA. It has half as many papers as the members of the second group combined: France, People's Republic of China, Germany, Japan, Russia, England, and Italy.

To determine the trends in countries containing the most Fractals papers, the results from 1991–1993 were examined. Table 5B summarizes results from the top 20 countries.

Table 5b. Prolific Countries (1991–1993).

Country	# of Papers
USA	1596
France	475
Germany	442
Japan	331
England	257
Italy	244
Canada	226
USSR	202
Peoples Republic of China	152
Israel	132
India	117
Russia	113
Spain	94
Netherlands	88
Switzerland	83
Poland	75
Australia	70
Norway	53
Denmark	48
Sweden	43
Brazil	40
Belgium	38
Greece	38
Scotland	35
Hungary	31
Argentina	30
Austria	29
Taiwan	27
Czechoslovakia	26
South Korea	25

The countries of the former Soviet Union had 337 papers in aggregate in 1991–1993, and 402 in aggregate in 2001–2002. The major shift is the increased ranking of People's Republic of China from ninth in 1991–1993 to third (or fourth, depending on whether the former Soviet Union is aggregated or not) in 2001–2002, and the concomitant increase in numbers of papers from 152 to 399.

Figure 6 contains a co-occurrence matrix of the top 15 countries for 2001–2002. In terms of absolute numbers of co-authored papers, the USA's major partners are France, Germany, Canada, England, Japan, and Italy. Interestingly, the USA is China's dominant major partner, having 2.5 times the number of co-authored papers with China (30) as China's next larger partner, Germany (12). Overall, countries in similar geographical regions tend to co-publish substantially, the USA being a moderate exception.

Figure 7 contains a co-occurrence matrix of the top 15 countries for 1991–1993. In terms of absolute numbers of co-authored papers, the USA's major partners are France, Germany, Israel, Italy, and Canada. Again, the USA was China's major partner, having slightly more co-authored papers with China (10) than China's next larger partners, Germany (8) and Italy (7).

Table 7. Country Co-Occurrence Matrix (1991–1993).

Country	Canada	England	France	Germany	India	Israel	Italy	Japan	Netherlands	Peoples R China	Russia	Spain	Switzerland	USA	USSR
Canada	226	3	25	4	2	3	2	3	2	1	0	3	0	32	3
England	3	257	8	6	4	1	7	4	3	2	3	7	2	25	3
France	25	8	475	23	0	12	23	2	10	0	3	5	10	79	5
Germany	4	6	23	442	1	15	11	10	5	8	7	1	19	54	5
India	2	4	0	1	117	0	0	2	0	0	0	0	0	7	0
Israel	3	1	12	15	0	132	3	0	1	1	6	2	9	44	3
Italy	2	7	23	11	0	3	244	1	4	7	4	0	5	34	2
Japan	3	4	2	10	2	0	1	331	4	4	1	4	1	26	0
Netherlands	2	3	10	5	0	1	4	4	88	3	1	1	1	17	0
Peoples R China	1	2	0	8	0	1	7	4	3	152	1	0	0	10	0
Russia	0	3	3	7	0	6	4	1	1	1	113	0	0	6	1
Spain	3	7	5	1	0	2	0	4	1	0	0	94	1	12	0
Switzerland	0	2	10	19	0	9	5	1	1	0	0	1	83	8	0
USA	32	25	79	54	7	44	34	26	17	10	6	12	8	1596	16
USSR	3	3	5	5	0	3	2	0	0	0	1	0	0	16	202

4.2 CITATION STATISTICS ON AUTHORS, PAPERS, AND JOURNALS

The second group of metrics presented is counts of citations to papers published by different entities. While citations are ordinarily used as impact or quality metrics (Garfield, 1985), much caution needs to be exercised in their frequency count interpretation, since authors cite or do not cite particular papers for numerous reasons (Kostoff, 1998b; MacRoberts and MacRoberts, 1996).

The citations in all the retrieved SCI papers were aggregated. The authors, specific papers, years, journals, and countries cited most frequently were identified, and were presented in order of decreasing frequency. A small percentage of any of these categories received large numbers of citations. From the citation year results, the most recent papers tended to be the most highly cited. The most recent papers reflect rapidly evolving fields of research.

4.2.1 Most Cited Authors

Table 8 lists the most highly cited authors from the 2001–2002 database. Many of these highly cited authors worked at various institutions throughout their careers, and the institution listed was their residence when some of the highly cited work was performed.

Table 8. Most Cited Authors (2001–2002)

(Cited By Other Papers In This Database Only),

Author	Institution	Country	# Cites
Mandelbrot B B	IBM	USA	1172
Bak P	Brookhaven National Lab	USA	614
Falconer K J	University of Bristol	UK	331
Meakin P	DuPont	USA	291
Tsallis C	CTR Brasileiro Pesquisas Fis	Brazil	290
Grassberger P	University of Wuppertal	Germany	221
Feder J	University of Oslo	Norway	203
Witten T A	Exxon Res & Eng	USA	187
Halsey T C	University of Chicago	USA	170
Frisch U	CNRS	France	158
Turcotte D I	Cornell University	USA	158
Vicsek T	Eotvos Lorand University	Hungary	157
Avnir D	Hebrew University	Israel	156
Metzler R	University of Ulm	Germany	146
Kolmogorov A N	Lomonosov State University	Russia	145
Stauffer D	KFA Julich Gmbh	Germany	144
Pfeifer P	University of Bielefeld	Germany	142
Elnaschie M S	Cornell University	USA	136
Benzi R	University of Rome Tor Vergata	Italy	131
Zaslavsky G M	Academy of Science USSR	Russia	128

Of the 20 most cited authors, 7 are from the USA, 8 from Western Europe, 3 from Eastern Europe, 1 from the Middle East, and 1 from Latin America. This distribution is far different from the most prolific authors of 2001–2002, where 8 of 19 were from the East/Far East. This distribution of most cited authors more closely resembles the distribution of most prolific authors from 1991–1993, where only one was from the Far East.

There are a number of potential reasons for this regional difference between most prolific and cited authors in 2001–2002 . The most prolific may not be the highest quality, or many of the most prolific authors could be relatively recent, and insufficient time has elapsed for their citations to accumulate. In another 3 or 4 years, when the papers from present-day authors have accumulated sufficient citations, firmer conclusions about quality can be drawn.

The lists of 19 most prolific authors from 2001–2002 and 20 most highly cited authors only had two names in common (ZASLAVSKY, TSALLIS). This phenomenon of minimal intersection has been observed in all other text mining studies performed by the first author. The lists of 18 most prolific authors from 1991–2093 and 20 most highly cited authors only had one name in common (MEAKIN). This disconnect is more disconcerting, since adequate time has accumulated in the past decade for these 1991–1993 papers to gather citations. A more detailed examination of all these papers would be required to resolve this dilemma, which is beyond the scope of this document.

Twelve of the most cited authors' institutions are universities, five are government-sponsored research laboratories, and three are private companies.

The citation data for authors and journals represent citations generated only by the specific records extracted from the SCI database for this study. It does not represent all the citations received by the references in those records; these references in the database records could have been cited additionally by papers in other technical disciplines.

4.2.2 Most Cited Papers

Table 9 lists the most highly cited documents from the 2001–2002 database.

Table 9. Most Cited Documents (Total Citations Listed in SCI).

Document	# Cites
Mandelbrot Bb, 1982, Fractal Geometry Nat	5107
Fractal Geometry of Nature	
Bak P, 1987, Phys Rev Lett, V59, P381	1731
Self-Organized Criticality	
Mandelbrot Bb, 1983, Fractal Geometry Nat	2942
Fractal Geometry of Nature	
Feder J, 1988, Fractals	2057
General Fractals	
Bak P, 1988, Phys Rev A, V38, P364	1279
Self-Organized Criticality	
Witten Ta, 1981, Phys Rev Lett, V47, P1400	2181
Diffusion-Limited Aggregation	
Halsey Tc, 1986, Phys Rev A, V33, P1141	1505
Fractal Measures and Their Singularities	
Mandelbrot Bb, 1968, Siam Rev, V10, P422	876
Fractional Brownian Motions and Noises	
Falconer K, 1990, Fractal Geometry Mat	415
Mathematical Foundations of Fractal Geometry	
Tsallis C, 1988, J Stat Phys, V52, P479	641
Generalization Of Boltzmann-Gibbs Statistics	
Vicsek T, 1992, Fractal Growth Pheno	478
Fractal Growth Phenomena	
Leland We, 1994, Ieee Acm T Network, V2, P1	371
Self-Similar Nature of Ethernet Traffic	
Barabasi Al, 1995, Fractal Concepts Sur	1026
Fractal Concepts in Surface Growth	
Havlin S, 1987, Adv Phys, V36, P695	918
Diffusion in Disordered Media	
Bouchaud Jp, 1990, Phys Rep, V195, P127	702
Anomalous Diffusion in Disordered Media	
Hentschel Hge, 1983, Physica D, V8, P435	920
Generalized Dimensions of Fractals and Strange Attractors	
Mandelbrot Bb, 1974, J Fluid Mech, V62, P331	686
Intermittent Turbulence in Self-Similar Cascades	
Hutchinson Je, 1981, Indiana U Math J, V30, P713	470
Fractals and Self Similarity	
MANDELBROT BB, 1984, NATURE, V308, P721	547
Fractal Character Of Fracture Surfaces Of Metals	
SAMORODNITSKY G, 1994, STABLE NONGAUSSIAN R	393
Stable Nongaussian Random Processes	

The theme of each paper is shown in italics on the line after the paper listing. The order of paper listings is number of citations by other papers in the extracted database analyzed. The total number of citations from the SCI paper listing, a more accurate measure of total impact, is shown in the last column on the right.

Physical Review Letters contains the most papers, two out of the 20 listed. A substantial number of books are listed (about one-third), noticeably larger than in other topics studied. Reasons for this difference are unclear.

Most of the journals are fundamental science journals, and most of the topics have a fundamental science theme. The majority of the papers are from the 1980s, with seven from the 1990s, and one paper from 1968.

Three Fractals books are in the top 20 cited documents. Several of the most cited papers are review articles. Otherwise, the most cited papers appear in physics journals focused on fractal motions, growth of fractal shapes, fractal noise, and fractal measures.

The list of most cited includes general books by Mandelbrot, and Feder, covering many fractals topics. Mandelbrot's book defined the field, and many papers refer to it. The paper of Bak is a theory called "self-organized criticality" of why natural objects can wind up as fractal shapes. The other themes cited are mostly **fractal motions** or fractal random processes (mostly generalizations on Brownian motion but with different scaling properties), or random walks called Levy flights with jump sizes on all scales. Another theme is **fractal noise**, i.e., fluctuations that are wild and fractal. A third theme is **fractal growth.** How can particles or clusters of particles aggregate into fractal shapes. How can fractal biological shapes, like the branching in the lung, grow, or how can shapes break down (dissolve, weather, etc.), leaving fractal shapes behind. A fourth theme is **fractal measures**. How can fractal objects be characterized? One way is with a fractal dimension. Another way is to treat the fractal dimension as a variable and get a distribution of fractal dimensions to describe fractal objects. Note that fractals are a condition that can arise within physical theories, to obtain fractal motions or fractal shapes under certain conditions.

Thus, the major intellectual emphasis of cutting-edge Fractals research, as evidenced by the most cited papers, is well aligned with the intellectual heritage and performance emphasis, as will be evidenced by the clustering approaches presented later.

4.2.3 Most Cited Journals

Table 10 lists the most highly cited journals from the 2001–2002 database.

Table 10. Most Cited Journals (Cited By Other Papers in This Database Only).

Journal	# of Cites
Phys Rev Lett	7048
Phys Rev E	3602
Astrophys J	3068
Phys Rev B	2395
Nature	1754
Phys Rev A	1609
Physica A	1335
J Fluid Mech	1208
J Phys A-Math Gen	1122
J Chem Phys	1061
Science	1001
Phys Rev D	992
Physica D	976
Mon Not R Astron Soc	875
Phys Lett A	851
J Colloid Interf Sci	847
Astron Astrophys	782
J Stat Phys	753
Phys Fluids	686
Water Resour Res	665

Three main groups of cited journals may be discerned. PHYS REV LETT received almost as many cites as the three journals in the next group (PHYS REV E, ASTROPHYS J, PHYS REV B), or even the first five journals in the following group (NATURE, PHYS REV A, PHYSICA A, J FLUID MECH, J PHYS A, J CHEM PHYS, SCIENCE). PHYS REV LETT emphasizes rapid publication of 'hot' topics, and would therefore tend to establish primacy in an emerging field. Since one aspect of citations is identifying the original literature of a new topic, a credible journal with these characteristics would tend to receive large numbers of citations.

Unlike the relatively disjoint relationship between most prolific authors in 2001–2002 and most cited authors in 2001–2002, the relationship between most prolific journals in 2001–2002 and most cited journals in 2001–2002 is much closer. Thirteen of the 20 most highly cited journals in 2001–2002 are also on the list of 19 most prolific journals in 2001–2002. The more applied journals on the most prolific list for 2001–2002 are replaced by the more fundamental journals on the most cited list for 2001–2002. Thirteen of the 20 most highly cited journals in 1991–1993 are also on the list of 20 most prolific journals in 1991–1993. All of the top 10 most prolific journals from 1991–1993 are on the list of 20 most highly cited journals of 2001-02. The more applied journals on the most prolific list for 1991–1993 are replaced by the more fundamental journals on the most cited list for 2001–2002.

The authors end this bibliometrics section by recommending that the reader interested in researching the topical field of interest would be well-advised to, first, obtain the highly cited papers listed and, second, peruse those sources that are highly cited and/or contain large numbers of recently published papers.

4.3 DT RESULTS

Two major analytic methods are used in this section to generate taxonomies of the SCI databases: concept clustering, based on phrase/word aggregation, and document clustering, based on document aggregation. Counting of documents within each major cluster provides some estimate of level of effort within the thematic area represented by the cluster.

4.3.1 Taxonomies

4.3.1.1 Concept Clustering. Two statistically based concept clustering methods were used to develop taxonomies: (1) factor matrix clustering, and (2) multi-link clustering. Both offer different perspectives on taxonomy category structure from the document clustering approach described later. None of the three approaches is inherently superior.

In this section, a synergistic combination of factor matrix and multi-link clustering is described that offers substantial improvement in the quality of the resultant clusters. Once the appropriate factor matrix has been generated, the factor matrix can then be used as a filter to identify the significant technical words for further analysis. Specifically, the factor matrix can complement a basic trivial word list (e.g., a list containing words that are trivial in almost all contexts, such as 'a', 'the', 'of', 'and', 'or', etc) to select context-dependent high technical content words for input to a clustering algorithm. The factor matrix pre-filtering will improve the cohesiveness of clustering by eliminating those words that are trivial words operationally in the application context.

The present application uses single words for clustering rather than the multi-word phrases of previous applications. While some of the technical detail is lost by excluding the ordering information contained in multi-word phrases, inclusion of all single words compensates for the elimination of multi-word phrases due to the selection algorithm of the Natural Language Processor.

4.3.1.1.1 Factor Matrix Clustering. In the factor matrix used, the rows are the words and the columns are the factors. The matrix elements Mij are the factor loadings, or the contribution of word i to the theme of factor j. The theme is determined by those words that have the largest values of factor-loading. Each factor had a positive value tail and negative value tail. For each factor, most of the time, one of the tails dominated in terms of absolute value magnitude. This dominant tail determined the central theme of each factor.

To generate the words input to the factor matrix, the highest frequency high technical content words were identified (819 words). A factor analysis was performed using the TechOASIS statistical package, and a factor matrix consisting of 30 factors resulted. A description of each factor, and the aggregation of all factors into a taxonomy, follows. The capitalized phrases in parentheses represent high factor-loading phrases for the factor described.

Factor 1 (Hausdorff, set, sets, infinity, hyperbolic, points, topological, dimension, infinite, Counting, Box) focuses on estimating the Hausdorff dimension and topological entropy of limit and hyperbolic sets, and relating Hausdorff dimension to Box-Counting dimension.

Factor 2 (microwave, cosmic, cosmological, background, dark, scale-invariant, matter, constraints, galaxy) focuses on anisotropies of cosmic microwave background observations and galaxy and

cluster surveys of large scale structure for theories of cosmic structure formation, to test the inflationary cold dark matter scenario of structure formation, and derive constraints on cosmological parameters.

Factor 3 (Monte, Carlo, percolation, lattice, site, lattices, sites, simulations, square, cluster, random, critical) focuses on Monte Carlo simulations of percolation processes in inhomogeneous lattices, emphasizing the influence of inhomogeneities on the parameters (critical concentration, average number of sites in finite clusters, percolation probability, critical exponents, and fractal dimension of an infinite cluster) characterizing the percolation in the system.

Factor 4 (landscape, spatial, forest, species, fragmentation, patterns, soil, pattern, areas, area, environmental, Population) focuses on patterns of landscape spatial structures, and impacts of changes in fragmented landscape patterns on habitats and populations of forest dwelling species.

Factor 5 (flow, velocity, turbulent, Reynolds, layer, fluid, shear, flows, turbulence, jet, self-similar, viscosity, viscous, mixing, pressure) focuses on Reynolds number dependency of self-similar structures in shear layers of turbulent viscous flows.

Factor 6 (microscopy, AFM, atomic, force, scanning, microscope, films, electron, SEM, film, thin, roughness, morphology, surface, substrate, deposition, deposited, images, substrates, nm, surfaces) focuses on determination by atomic force and scanning electron microscopies of surface topography of thin films deposited on surfaces.

Factor 7 (chaotic, bifurcation, Poincare, Lyapunov, periodic, attractors, chaos, attractor, map, dynamical, orbits, attraction, basin, dynamics, unstable, oscillations, feedback, maps, mapping) focuses on chaotic motions of dynamical systems, using Poincare maps and associated bifurcation diagrams to display chaotic attractors and Lyapunov direct method to determine equilibria stability, with emphasis on basins of attraction of chaotic attractors.

Factor 8 (Brownian, fractional, motion, FBM, noise, Hurst, Gaussian, motions, stochastic, white) focuses on use of fractional Brownian motion, parameterized by the Hurst exponent, and white noise as models in time series analysis, to estimate stochastic properties of time series with fractal noise behavior.

Factor 9 (crack, stress, fracture, plastic, strain, deformation, tip, loading, elastic, stresses, material, mechanical, materials, specimens) focuses on fractal nature of fractal surfaces, emphasizing the relation of crack growth to stress and strain fields, especially in elastic materials with crack-tip plastic deformation assumptions.

Factor 10 (self-organized, criticality, SOC, avalanches, avalanche, sandpile, solar, critical, cellular, state, automata, plasma, activity) focuses on self-organized critical systems exhibiting power law properties, mainly the avalanches of events described by sand-pile models, with some emphasis on geometrical properties of avalanches in self-organized critical models of solar flares and their impact on the magnetosphere through the solar wind.

Factor 11 (star, accretion, gas, galaxies, disk, cloud, emission, galaxy, mass, density, wind) focuses on gas accretion in line-emission disk galaxies, emphasizing low-mass star formation from cloud core condensation.

Factor 12 (turbulence, scaling, multi-fractal, scales, turbulent, intermittency, exponents, intermittent, scale, inertial, dissipation, fluctuations, velocity) focuses on multi-fractal descriptions of fine-scale turbulence structure.

Factor 13 (patients, heart, age, bone, controls, groups, blood, fractures, group) focuses on heart-rate dynamics of patients age-matched with control groups, as well as comparing bone architecture in patients with osteoporotic fractures and in controls matched on bone mass.

Factor 14 (pore, porous, pores, porosity, permeability, media, fractal, medium, adsorption, water) focuses on relation of permeability to fractal dimensions for porous media.

Factor 15 (wave, waves, propagation, shock, nonlinear, equations, front, equation, heat, medium, differential, media) focuses on propagation of shock waves and fronts in gaseous media, and the nonlinear differential equations used to describe the flow.

Factor 16 (anomalous, Levy, diffusion, Fokker-Planck, exponential, random, long, walks, distributions, walk, equation, distribution) – focuses on anomalous diffusion from Fokker-Planck particle distribution equations driven by Levy stable noise

Factor 17 (earthquake, seismic, earthquakes, fault, zones, zone, major, event, active) focuses on earthquake hazard assessment, and prediction of major events from precursor spatial and temporal seismicity patterns in active fault zones.

Factor 18 (gel, gels, gelation, polymer, colloidal, aqueous, concentration, protein, polymers, chains, pH, molecules, dynamic, aggregation) focuses on colloidal and polymer gels, especially based on aqueous solvents for the sol-gel reaction, emphasizing the role of suspension anion concentration to promote rapid aggregation.

Factor 19 (aggregates, particles, aggregate, aggregation, particle, size, coagulation, light, scattering, diameter, colloidal, clusters, primary, nm, sizes, D-f, cluster) focuses on aggregation kinetics of colloidal suspensions of particles with varying sizes, emphasizing dynamic light scattering to measure particle diameters, and predicting aggregation rate and critical coagulation concentration.

Factor 20 (power, frequency, frequencies, spectral, dielectric, spectra, law, slope, noise, spectrum, peak) focuses on power law modeling of frequency-dependent dielectric spectra.

Factor 21 (traffic, network, packet, networks, bandwidth, self-similarity, control, long-range) focuses on self-similar traffic in high-bandwidth packet communication networks.

Factor 22 (image, images, coding, feature, recognition, algorithm, compression, neural, wavelet, blocks, transform, texture, features, algorithms) focuses on texture feature coding for image classification, using encoding algorithms containing feature extraction and recognition for image compression.

Factor 23 (X-ray, scattering, small-angle, neutron, diffraction, angle, silica, nm, pore, pores, gel, electron, intensity, crystal, microscopy) focuses on use of small-angle X-ray scattering, electron microscopy, and neutron scattering to measure pore sizes, particle size, and surface roughness, especially on gels, silica particles, and crystals.

Factor 24 (growth, island, islands, aggregation, deposition, morphology, Carlo, Monte, kinetic, nucleation, morphologies, deposited, diffusion-limited, substrate, temperatures, formation, kinetics, step, cluster, substrates) focuses on growth of films (as a function of temperature) by deposition, diffusion, and aggregation (DDA) on percolation substrates, and parallel kinetic modeling using Monte Carlo techniques.

Factor 25 (hole, black, gravitational, collapse, gravity, symmetric, singularity, accretion, momentum) focuses on critical phenomena in gravitational collapse, and the shared features with the dynamics of singularity or black hole formation (universality, self-similarity, scaling).

Factor 26 (rough, roughness, surfaces, surface, self-affine, Carlo, Monte, scattered, scattering, angle, fractal) focuses on scattering from self-affine rough surfaces, emphasizing Monte Carlo simulation of rough surface scattering, and modeling of the surfaces as fractal.

Factor 27 (adsorption, rate, energy, rates, reaction, kinetic) focuses on measurement of fractal surfaces and associated energy distributions based on absorption kinetics, and relation of fractal surfaces to enhanced reaction rates.

Factor 28 (pattern, structural, fractures, patterns, bone) focuses on fracture patterns in bones.

Factor 29 (magnetic, field, electrons, electric, plasma, localized, current, electron, anomalous, quantum, ion, fields) focuses on quantum nature of anomalous electron diffusion, and associated electron currents, in plasmas in a magnetic field.

Factor 30 (phase, canonical, thermodynamic, equilibrium, transitions, transition, phases, ensemble, critical, temperature) focuses on phase transitions and relaxation to thermodynamic equilibrium of canonical ensembles.

(In the next section, a taxonomy is generated using the multi-link hierarchical clustering approach. The 30 factors above are assigned to the appropriate categories in the taxonomy, providing good coverage and an excellent match.)

After the 30-factor matrix was generated, it was then used for word filtering and selection. In the present study, the 819 words in the factor matrix had to be culled to the approximately 250 allowed by the Microsoft Excel-based clustering package, WINSTAT. The 250-word limit is an artifact of Excel. Other software packages may allow more or less words to be used for clustering, but all approaches perform culling to reduce dimensionality. The filtering process presented here is applicable to any level of filtered words desired.

The factor loadings in the factor matrix were converted to absolute values. Then, a simple algorithm was used to automatically extract those high factor loading words at the tail of each factor. If word variants were on this list (e.g., singles and plurals), and their factor loadings were reasonably close (12), they were conflated (e.g., 'agent' and 'agents' were conflated into 'agents', and their frequencies were added). A few words were eliminated manually, based on factor loading and estimate of technical content.

4.3.1.1.2 Multi-Link Clustering. A symmetrical co-occurrence matrix of the 253 highest frequency, high technical content words was generated. The matrix elements were normalized using the Equivalence Index (Eij=Cij^2/Ci*Cj, where Ci is the total occurrence frequency of the ith phrase,

and Cj is the total occurrence frequency of the jth phrase, for the matrix element ij), and a multi-link clustering analysis was performed using the WINSTAT statistical package. The Average Linkage hierarchical aggregation method was used. A description of the final 253 phrase dendogram (a hierarchical tree-like structure), and the aggregation of its branches into a taxonomy of categories, follows. Figure 1 is the dendogram of the 253 words. One axis is the words, and the other axis ('distance') reflects their similarity. The lower the value of 'distance' at which words, or word groups, are linked together, the closer their relation. As an extreme case of illustration, words that tend to appear as members of multi-word phrases, such as 'fractional Brownian motions', 'Monte Carlo', or 'self-organized criticality', appear adjacent on the dendogram with very low values of 'distance' at their juncture. The capitalized phrases in parentheses represent cluster boundary phrases for each category.

The 253 phrases in the dendogram are grouped into 28 elemental clusters. These clusters form the lowest level of the taxonomy hierarchy. Each cluster is assigned a letter, ranging from A to AB. The cluster hierarchies are determined by the branch structure of the dendogram. Two main braches (clusters) are at the highest hierarchical level. Starting from the phrase adjoining the 'distance' ordinate, the first main cluster (A-T) ranges from FRACTAL to AUTOMATA. The second main cluster (U-AB) ranges from SURFACES to MAJOR, and is moderately smaller in extent than the first main cluster. While the total dendogram reflects different aspects of Fractals, the first cluster (A-T) covers Fractals in the dynamical systems context, while the second cluster (U-AB) covers Fractals in static structures. Each of these highest level clusters will be divided and sub-divided into smaller clusters, and discussed.

Cluster (A-T) can be divided into clusters (A-S) and (T). Cluster (A-S) ranges from FRACTAL to PEAK, and cluster (T) ranges from SPECIES to AUTOMATA. Cluster (A-S) focuses on dynamical systems aspects of Fractals for physical, engineering, and life sciences, while cluster (T) focuses on environmental ecosystems, emphasizing estimation of equilibrium populations of species inhabiting fragmented landscapes and subject to fragmented resources.

Cluster (U-AB) can be divided into clusters (U-Z) and (AA-AB), where cluster (AA-AB) is much smaller than cluster (U-Z). Cluster (U-Z) ranges from SURFACES to PROTEIN, and cluster (AA-AB) ranges from MATERIALS to MAJOR. Cluster (U-Z) focuses on surface topology at micro-scales, while cluster (AA-AB) focuses on continuum mechanics of materials at macro-scales.

Before the elemental clusters are described, the meta-level description above needs to be brought into a larger perspective. Most previous text mining studies performed by the first author focused on technical disciplines. These disciplines ranged from relatively focused (e.g., hypersonic flow or fullerenes) to relatively broad (e.g., aircraft or chemistry). A strong disciplinary thread throughout the data allows division of thematic categories into relatively crisp and complementary sub-categories. At any hierarchical level in the taxonomy, the categories are sharply defined and complementary (e.g., aircraft could sub-divide into fixed wing or movable wing, or converters could divide into direct electrical converters and thermal step converters).

For those few text mining studies that did not focus on a discipline directly, but focused on applications of a discipline (such as papers that cite a discipline or patents that cite a discipline), the taxonomy categories have a different type of structure. In those cases, the discipline thread that links the categories is weaker, and there is a competition in the algorithm between application sub-division and thematic sub-division. Sometimes the thematic sub-division will dominate, and sometimes the application sub-division will dominate.

Fractals is a unique type of subject area. It is not a discipline in the sense of chemistry or physics, but rather is a characteristic feature or property of a system or process. Many fractals papers focus on the application, and only a relatively modest number address the intrinsic nature of fractals. The consequence for clustering is that the cluster themes are not thematically pure in every case. In particular, while the top-level categorical division into dynamics (essentially time series analysis) and statics (essentially spatial pattern analysis) holds in the large, in a few sub-categories the application theme is stronger than the top level fractals division, and the elemental cluster theme will reflect the full application theme rather than a dynamic or static component of the applications theme. In the descriptions of the elemental clusters that follow, these few anomolous cases will be identified.

Cluster A (FRACTAL to AREAS) focuses on fractal dimensions, based on size distribution functions, to characterize spatial patterns over scaling ranges.

Cluster B (SPECTRA to HEAT) focuses on multi-fractal analysis of power law fluctuation spectra.

Cluster C (CRITICAL to ACTIVITY) focuses on avalanche properties of sandpile models that exhibit self-organized criticality.

Cluster D (FIELDS to ION) focuses on magnetic and electric fields, especially the solar wind plasma currents.

Cluster E (SIMULATIONS to SIERPINSKI) focuses on Monte Carlo simulations of square lattices, emphasizing Ising spins located at the sites of Sierpinski carpets.

Cluster F (EQUATIONS to EXPONENTIAL) focuses on nonlinear differential equations, especially Fokker-Planck, and addresses random walk models of anomalous diffusion with Levy distributions.

Cluster G (MOTIONS to WHITE) focuses on the self-similar Gaussian process of fractional Brownian motion that includes stochastic white noise and a Hurst parameter to explain the complexity.

Cluster H (DYNAMICS to OSCILLATIONS) focuses on chaotic dynamics, using Lyapunov exponents to predict chaotic behavior and Poincare maps to display chaotic attractors. The fractal structure of the basins of attraction and the orbits of the period-multiplying bifurcations are emphasized.

Cluster I (SETS to CANONICAL) focuses on Hausdorff dimension and topological entropy of hyperbolic sets.

Cluster J (RATES to FEEDBACK) focuses on heart rate variability of patients in different age-matched control groups, with associated blood pressure monitoring.

Cluster K (FEATURES to TRANSFORM) focuses on pattern recognition of textured images using fractal features, with emphasis on bone texture roughness and anisotropy. This cluster has static and dynamic components. The origin can be seen more clearly by examining factor 13, a factor that incorporates the medical components of clusters J and K. In factor 13, heart rate dynamics and bone

architecture are combined under one medical theme. The medical application has more influence on the factor theme than the division into its dynamic and static components.

Cluster L (STATISTICAL to NEURAL) focuses on statistical models of long-range dependent network traffic, including the self-similarity in packet network traffic.

Cluster M (SLOPE to DIELECTRIC) focuses on local Box-Counting method for computing scale-dependent (local) fractal dimensions.

Cluster N (SELF-SIMILARITY to WALL) focuses on self-similar solutions that describe the dynamics of turbulent viscous fluid flows, assuming turbulent energy dissipation to be a power law of the density and velocity, and mass is conserved.

Cluster O (MOMENTUM to HEATING) focuses on accretion disks with jets, especially heating of the accretion disks by energy radiated from the infalling material.

Cluster P (WAVES to FRONT) focuses on shock wave and front propagation in gas.

Cluster Q (LINE to CASCADE) focuses on line emissions from stars and associated clouds, including information provided about the cloud's core.

Cluster R (QUANTUM to SINGULARITY) focuses on quantum self-similar gravitational collapses and black hole formation.

Cluster S (MATTER to PEAK) focuses on anisotropies of cosmic microwave background observations and galaxy and cluster surveys of large-scale structure for theories of cosmic structure formation, to test the inflationary cold dark matter scenario of structure formation, and derive constraints on cosmological parameters.

Cluster T (SPECIES to AUTOMATA) focuses on forest-dwelling species' population dynamics, especially in heterogeneous fragmented landscape environments, emphasizing hierarchical cellular automata modeling. This cluster has static and dynamic components. The reasons for the dichotomy can be seen somewhat more transparently by examining its associated factor 4. The split into population dynamics and landscape patterns is dominated by the applications theme of environmental dynamics. Interestingly, on the dendogram, this cluster is positioned at the juncture between the dynamics and statics highest level categorization, reflecting its association with each category.

Cluster U (SURFACES to ISLANDS) focuses on determination by atomic force and scanning electron microscopies of self-affine roughness topography of thin films deposited on surfaces.

Cluster V (PARTICLES to REACTION) focuses on diffusion-limited aggregation of particles into colloidal percolation clusters.

Cluster W (SCATTERING to DIFFRACTION) focuses on scattering, emphasizing light scattering, small-angle X-ray scattering, and neutron scattering.

Cluster X (CONCENTRATION to D-f) focuses on colloidal and polymer gels, especially based on aqueous solvents for the sol-gel reaction, emphasizing the role of suspension anion concentration to promote rapid aggregation.

Cluster Y (MEDIA to HETEROGENEITY) focuses on relation of permeability to fractal dimensions for porous media.

Cluster Z (POLYMERS to PROTEIN) focuses on protein and polymer chains, emphasizing the fractal character of protein and polymer molecules.

Cluster AA (MATERIALS to SPECIMENS) focuses on material fracture, emphasizing the relation of crack growth to stress and strain fields, especially in elastic materials with crack-tip plastic deformation assumptions.

Cluster AB (ZONES to MAJOR) focuses on seismic activity in fault zones, and relation to major earthquake events. This cluster has static and dynamic components. The reasons for the dichotomy can be seen somewhat more transparently by examining its associated factor 17. The split into seismic signal dynamics for earthquake analysis and fault zone patterns for earthquake prediction is dominated by the applications theme of earthquakes.

Table 11 shows the assignment of factors and individual clusters to the second-level categories of the multi-link clustering-defined taxonomy. Correspondence between the factors and categories is good.

Table 11. Factor Matrix-Word Cluster Taxonomy.

FRACTALS TAXONOMY

Microscale		STATIC Macroscale		Phys/Eng/ Life Sci		Dynamic Envir Sci	
CL WRD	FAC	CL WRD	FAC	CL WRD	FAC	CL WRD	FAC
U	6, 24	AA	9	A	1	T	4
V	19, 27	AB	17	B	20		
W	23, 26			C	10		
X	18			D	29		
Y	14			E	3		
Z	18			F	16		
				G	8		
				H	7		
				I	1, 30		
				J	13		
				K	28		
				L	21		
				M	1		
				N	5, 12		
				O	11, 25		
				P	15		
				Q	11		
				R	25		
				S	2		

4.3.1.1 Document Clustering. Document clustering is the grouping of similar documents into thematic categories. Different approaches exist (e.g., Willett, 1988; Rasmussen, 1992; Cutting et al., 1992; Guha, Rastogi, and Shim, 1998; Hearst, 1998; Zamir and Etzioni, 1998; Karypis, Han, and Kumar, 1999; Steinbach, Karypis, and Kumar, 2000). The approach presented in this document is based on a Greedy String Tiling (GST) text-matching algorithm (Wise, 1992, Prechelt, Malpohl, and Philippsen, 2002). Because this document is the first time that GST text clustering has been published, it is described in some detail in Appendix A. Basically, GST clustering forms groups of documents based on the cumulative sum of shared strings of words. Each group is termed a cluster, and the number of records in each cluster, and the highest frequency technical keywords in each cluster, are two outputs central to this analysis.

The 64 clusters with the largest number of Abstracts were extracted, and are listed below. The main keywords from each cluster are shown in parentheses after the cluster number, and the number of records in each cluster is shown in parenthesis before the cluster number. The keywords are arranged in frequency of appearance, in descending order. Three levels of filtering were used to obtain the main keywords shown below. First, a trivial word list (e.g., of, the, on, etc.) was applied to the raw data. Second, only the 30 highest frequency words for each cluster were retained. Third, a manual filtering was performed on the 30 highest words. Because of space limitations, the theme of

each cluster will not be written. The themes of each cluster are defined by the keywords shown. The taxonomy based on these themes follows the theme keyword listings.

(348) Cluster 1 (fractal, time, model, self, structure, scaling, distribution, multifractal, system, space, function, similar, dynamics, properties, dimension, systems, phase, scale)

(315) Cluster 2 (alpha, dimension, fractal, similar, time, infinity, percolation, random, scaling, set)

(306) Cluster 3 (fractal, dimension, surface, surfaces, model, fracture, distribution, roughness, structure, size, dimensions, scale, dimensional, function, adsorption, paper, area, pore, self, properties)

(186) Cluster 4 (fractal, scattering, aggregates, particles, size, dimension, structure, particle, small, aggregation, similar, light, mass, surface, angle, model, measurements, concentration, nm, aggregate, range, clusters)

(170) Cluster 5 (self, similar, solutions, solution, time, field, model, equation, equations, critical, similarity, flow, collapse, density, wave, shock, nonlinear, system, energy, evolution, one, initial, state, boundary, dimensional, function, numerical, grain, dynamics)

(129) Cluster 6 (flow, velocity, turbulent, similar, self, jet, layer, flame, turbulence, numerical, scale, model, pressure, boundary, time, surface, field, mean, range, density, experimental, structure, scaling, rate, conditions)

(121) Cluster 7 (power, model, law, distribution, self, organized, distributions, SOC, system, critical, time, scale, criticality, size, models, dynamics, field, statistical, avalanche, state, systems, avalanches, magnetic, fluctuations, transport, solar)

(86) Cluster 8 (chaotic, fractal, system, dynamics, set, dimension, systems, periodic, model, dynamical, Lyapunov, phase, scattering, attractors, orbits, structure, attractor, control, invariant, conditions, chaos, time, motion, parameter, numerical, space, initial, saddle, map)

Cluster 9 (blank Abstracts)

(81) Cluster 10 (image, fractal, images, feature, features, coding, algorithm, texture, scale, recognition, blocks, compression, domain, color, set, invariant, classification, wavelet, encoding, segmentation, dimension, pattern, information, transformation, time)

(79) Cluster 11 (similar, density, mass, gas, self, model, power, emission, distribution, luminosity, star, temperature, profile, ray, accretion, high, models, galaxies, structure, disk, fractal, regions, line, clusters, time, formation)

(77) Cluster 12 (fractal, patients, bone, hr, heart, variability, trabecular, dimension, rate, images, scaling, time, short, mean, frequency, ventricular, dynamics, atrial, measures, correlation)

(75) Cluster 13 (diffusion, anomalous, equation, time, fractional, equations, model, processes, Fokker, Planck, process, nonlinear, random, solutions, transport, field, law, solution, distribution, fractal, models, relaxation, Levy)

(71) Cluster 14 (surface, fractal, roughness, AFM, dimension, microscopy, films, atomic, force, deposition, surfaces, diffusion, morphology, time, growth, film, adsorption, concentration, dimensions, temperature, layer, scaling, polymer, structure, images)

(71) Cluster 15 (aggregation, model, particles, diffusion, limited, fractal, cluster, aggregates, clusters, growth, reaction, particle, size, rate, kinetics, dimension, surface, DLA, time, process, lattice, structure, phase, magnetic, dimensional, colloidal)

(67) Cluster 16 (dimension, set, Hausdorff, sets, self, Julia, similar, condition, boundary, function, systems, Box, class, Counting, hyperbolic, fractal, dimensional, attractor, dimensions)

(66) Cluster 17 (Brownian, fractional, motion, random, time, FBM, dimension, self, functions, process, stochastic, Gaussian, processes, noise, wavelet, model, order, stationary, similar, correlation, function, models, Hurst)

(64) Cluster 18 (critical, model, exponents, transition, fractal, percolation, simulations, models, system, scaling, properties, point, random, dimensions, dimension, phase, size, field, order, dimensional, lattice, multifractal, self, distribution, systems, Monte, Carlo, finite, interface, universality)

(60) Cluster 19 (traffic, network, model, self, time, similarity, similar, range, packet, paper, data, networks, cell, long, distribution, bandwidth, control, processes, process, wavelet, buffer, parameters, models, measurements, burst, probability, scheme, queueing, loss, fractional)

(55) Cluster 20 (time, fractal, series, data, scaling, rainfall, process, properties, power, dimension, long, dynamics, range, scale, model, correlations, scales, system, law, space, processes, self, statistical, structure, correlation, frequency, parameters, state, exponent, multifractal)

(52) Cluster 21 (growth, fractal, diffusion, surface, island, model, patterns, ion, temperature, deposition, formation, islands, high, processes, flux, edge, cm, compact, adatoms, pattern, morphology, shape, step, limited, aggregates, aggregation, mechanism)

(50) Cluster 22 (magnetic, field, current, model, solar, fractal, wind, energy, turbulence, plasma, reconnection, structures, scale, flux, sheet, self, system, structure, similar, fluctuations, dimensional, properties, systems, models, intermittency, observations, anomalous, intermittent, conductance, phase)

(42) Cluster 23 (scale, spectrum, models, power, omega, invariant, density, fluctuations, inflation, model, phi, large, cosmological, CMB, similar, background, universe, cosmic, equal, mass, matter, structure, microwave)

(39) Cluster 24 (dielectric, temperature, fractal, frequency, field, model, relaxation, spin, dependence, low, response, magnetic, power, solid, law, proton, phase, reaction, surface, time, high, similar, range, dimension)

(37) Cluster 25 (size, distribution, fractal, particle, dimension, distributions, self, coagulation, clusters, particles, number, model, rate, cluster, similar, large, membranes, gel, temperature, sizes, scaling, process, law, theory, step, experimental)

(35) Cluster 26 (habitat, landscape, spatial, landscapes, model, fragmentation, pattern, land, forest, area, fractal, threshold, patterns, population, cover, loss, models, dispersal, extinction, structure, persistence, resolution, patch, dimension, size, mean, ecological, fire, time, process)

(30) Cluster 27 (earthquakes, earthquake, seismic, model, large, time, fault, law, stress, seismicity, area, value, Gutenberg, fractal, release, critical, magnitude, richter, event, correlation, km, prediction, self, region, activity, spatial, system, scaling, models, energy)

(30) Cluster 28 (multifractal, dimensions, measures, spectrum, measure, alpha, spectra, local, dimension, properties, random, sets, dimensional, field, phase, scaling, sequences, transition, distribution, length, function, curve, dynamical, order, self, nature)

(28) Cluster 29 (gel, gelation, gels, model, fractal, time, point, structure, transition, sol, properties, omega, phase, measurements, alpha, temperature, growth, similar, behavior, experimental, system, systems, frequency, viscoelastic, network, range, concentration)

(24) Cluster 30 (phase, space, eta, fractal, range, order, model, set, distribution, similar, dimensions, self, system, energy, time, dimension, first, field, law, theory, dimensional)

(24) Cluster 31 (fractal, functions, interpolation, systems, curves, model, geometry, curve, solutions, generate, wavelet, function, dimensional, phi, dimension, dimensions, representation, objects, tool, approximation, standard, structure)

(24) Cluster 32 (porous, fractal, media, diffusion, pore, particles, column, fluid, model, dynamic, high, adsorbent, system, length, medium, coefficient, surface, vf, adsorptive, size, scale, ratio, viscous, heterogeneity, diameter, phase, profiles)

(22) Cluster 33 (function, time, dynamics, fractal, distribution, correlation, scaling, dimension, power, local, field, phase, law, equation, exponent, dimensional, long, fluctuations, simulations, exponents, models, system, probability, dynamic, length, agreement, experimental, model, space)

(22) Cluster 34 (existence, solutions, equations, fractal, global, partial, finite, attractor, domains, boundary, dimension, nonlinear, initial, attractors, type, equation, conditions, estimate, bounded, terms, time, order, system, class, solution, omega, Hausdorff)

(22) Cluster 35 (wave, laser, wind, time, plasma, dynamics, equations, systems, diffusion, order, particles, distribution, scale, waves, temperature, process, phase, self, model, law, power, exponent)

(21) Cluster 36 (scale, spatial, scales, patterns, species, small, size, structure, large, area, range, variation, similar, km, fractal, density, complexity, scaling, areas, soil, regions, steady, reservoir, pattern)

(21) Cluster 37 (data, model, models, sets, fractal, velocity, space, information, set, basalt, multiple, algorithm, scale, stochastic, real, seismic, finite, synthetic, vectors, dimension, structure)

(21) Cluster 38 (quantum, classical, system, chaotic, dynamics, diffusion, systems, localization, wave, dynamical, time, regular, statistical, anomalous, atoms, space, periodic, accelerator, phase, probability, particle, electron, energy, particular, potential, noise, chaos, mixed, kicked, fractal, structures, fluctuations)

(20) Cluster 39 (optical, fractal, field, scattering, light, surfaces, surface, distributions, intensity, properties, silver, experimental, clusters, rough, absorption, wavelength, photon, plasmon, numerical, local, scattered, modes, particles, dimensional, fields, aggregates, roughness, enhancement, microscopy)

(19) Cluster 40 (crack, stress, fracture, growth, tip, fractal, self, thickness, similar, model, bond, loading, cracks, mode, fatigue, strain, rate, small, material, factor, plastic, process, intensity, scale, length, conditions, cyclic, static, yielding)

(19) Cluster 41 (time, transport, properties, ray, chaotic, particle, flow, diffusion, anomalous, vortex, particles, motion, distribution, travel, tracers, exponent, tracer, long, system, space, velocity, vortices, function, stochastic, resonant, chaos, cores, phase, frequencies, point, region)

(18) Cluster 42 (fractal, dimension, consonants, particles, time, Box, domain, method, line, correlation, dynamics, signals, measurement, methods, structure, group, long, information, points, Counting, patterns, temporal, movement, domains)

(18) Cluster 43 (noise, signal, spectra, spatial, power, method, temporal, fractal, model, spectrum, process, low, system, stochastic, frequency, random, self, frequencies, filter, similar, correlation, function, range, film)

(17) Cluster 44 (scaling, multifractal, cascade, model, surface, scale, data, turbulence, random, models, multiplicative, scales, strong, field, drop, intermittency, anisotropic, extensive, similar, statistical, range, realistic, corresponding, parameter, layer, statistics, exponents, energy)

(17) Cluster 45 (relaxation, time, glass, phase, exponential, space, fractal, systems, stretched, energy, temperature, random, decay, spin, system, power, exponent, walks, percolation, law, triplet, transition, constant, ising, ergodic)

(17) Cluster 46 (dimension, fractal, function, Brownian, motion, model, surface, processes, graph, random, time, set, Hausdorff, process, network, properties, integral, transport, fractures, SSCC)

(16) Cluster 47 (fractal, cells, dimension, dendritic, complexity, area, cell, morphological, areas, whale, morphology, cancer, invasion, neuronal, mink, dimensions, visual, fin, mlt, pattern, low, determination, branching, quantitative, quantify, cortical, kinase, process)

(16) Cluster 48 (theory, physics, system, new, self, quantum, physical, statistics, biology, percolation, model, processes, fractal, critical, turbulence, chemical, properties, density, field, solution, modeling, aging, path, biological, consciousness, type, entropy, dynamics, distribution, systems)

(16) Cluster 49 (strain, dislocation, rate, stress, flow, plastic, density, fractal, deformation, statistical, critical, size, correlation, dynamics, multifractal, dimension, distributions, time, self, stage, fluctuation, specimens, load, high, large, surface)

(16) Cluster 50 (fault, shear, faults, fluid, zones, fractal, model, deformation, strain, system, zone, structure, systems, structural, structures, evolution, active, fold, time, ore, brittle, wedge, growth, scale, flow, mechanical, displacement, similar, distribution, properties, features)

(16) Cluster 51 (fractal, domains, X, ray, phase, temperature, microscopy, surface, structure, transition, film, domain, force, water, formed, high, nm, interface, condensed, monolayers, atomic, growth, degreesC, miscut, ripples, magnetic, air, scanning, diffraction, membranes, pressure, NCER)

(15) Cluster 52 (species, abundance, community, fractal, size, distribution, area, spatial, soil, range, diversity, structure, patterns, log, sample, tunnels, communities, termites, termite, samples, taxon, protozoan, number, form, similarity, scales, body, habitat, flavipes, dimension)

(15) Cluster 53 (wavelet, scale, singularities, fractal, function, functions, exponents, similar, self, transform, based, structures, multi, dimension, distribution, similarity, seismic, holder, modulus, maxima, structure, singularity, series, local, turbulence, multifractal, estimates, random)

(15) Cluster 54 (scaling, river, networks, stream, network, area, basins, drainage, basin, flow, model, law, rainfall, structure, channel, water, random, self, properties, models, distributions, function, spatial, statistical, ratio, fluctuations, slope, density, series, scales, resolution)

(15) Cluster 55 (antenna, fractal, antennas, Sierpinski, size, small, microstrip, patch, radiation, Koch, resonant, design, input, properties, frequency, curve, Carpet, multiband, pattern, structure, experimental, plane, similar, smaller, applications, square, geometry, conventional, novel)

(14) Cluster 56 (flow, fractal, network, blood, model, transport, heterogeneity, dimension, equations, pressure, vascular, aa, av, system, va, anastomoses, realizations, tests, lung, ttts, time, size, range, simulations, fracture, neural, vessel)

(13) Cluster 57 (temperature, transition, superfluid, measurements, magnetization, fractal, experimental, order, critical, glasses, range, field, elastic, pressure, disorder, structure, aerogel, heat, glass, canted, phase, substrate, reversal, atiferromagnetic, ferrimagnetic, bulk, magnetic)

(13) Cluster 58 (fractal, Ge, growth, films, electron, transmission, microscopy, shape, crystallization, bilayer, structure, Au, formation, nucleation, random, experimental, morphology, interface, networks, situ, time, model, mechanism, microscope, branching, rheological, formed, ratio, crystalline, amorphous, Pd)

(12) Cluster 59 (model, system, fractal, parameters, financial, dynamics, chaotic, fish, sediment, objects, frequency, conditions, uv, communication, mathematical, contour, plankton, experimental, scale, dynamic, trajectories, macro, radiation, rigid, complicated, economic, mixed, active, Hopf)

(12) Cluster 60 (spatial, soil, moisture, structure, scale, variability, scaling, scales, temporal, distribution, patterns, grassland, dynamics, fields, heterogeneity, disturbed, dye, properties, statistical, resolution, disturbance, elements, models, field, processes, lacunarity, landscape, random, large, areas, time)

(12) Cluster 61 (model, distribution, metastases, function, statistics, scale, permittivity, mean, variance, relationship, dispersion, flow, clustering, similar, poisson, image, blood, images, size, correlation, number, self, invariant, power, natural, linear, predictions, tumor, organ)

(12) Cluster 62 (self, fractal, scaling, affine, strain, dimension, similar, exponents, stress, structure, scale, exponent, roughness, collapse, experimental, interactions, biological, magnetic, phase, images, dielectric, statistical, range, line, simulations, universal, sample, measure, curve)

(12) Cluster 63 (systems, time, series, theory, interaction, test, structure, phase, interacting, dynamical, synchronization, correlation, dynamics, nonlinear, FMRI, function, self, model, space, Levy, dimension, complex, fractal, system, rate, turbulent, species)

(12) Cluster 64 (fractal, fractals, structure, Raman, properties, scattering, frequency, geometry, truncated, antennas, similar, intensity, self, space, mechanical, mathematical, staircase, nanocracks, process, low, nature, material, natural, shapes, biological, hot, vibrations, dimension, complex)

The taxonomy defined by the word-clustering algorithm was modified to include all the clusters in the document clustering. Specifically, a generic category was added to the second-level static and dynamic categories. Each cluster was assigned to the most appropriate category in the modified taxonomy defined by the WINSTAT-generated dendogram of the last section, based on the theme suggested by the highest frequency technical keywords. The number of records in each taxonomy category from all the clusters in the category was calculated, and is shown in Table 12. In this table, the top two levels of the taxonomy are presented. The top hierarchical level is composed of STATIC and DYNAMIC categories, and the second hierarchical level is composed of GENERIC STATIC, MICROSCALE, MACROSCALE, GENERIC DYNAMIC, PHYSICAL/ENGINEERING/LIFE SCIENCES, ENVIRONMENTAL SCIENCES. The first column is the cluster number, and the matrix elements are the number of records in the cluster in the specific second-level taxonomy category. The numbers in each second-level category are summed, and are summed in turn to give the total number of documents in each of the two first-level categories.

Table 12. Document Clustering Taxonomy.

Clust#	Generic	Phys/ Eng/Life	Dynamic Envir	Generic	Micro-scale	Static Macro-scale
1	348					
2				315		
3					306	
4					186	
5		179				
6		129				
7	121					
8		86				
9						
10		81				
11		79				
12		77				
13		75				
14					71	
15					71	
16	67					
17		66				
18	64					
19		60				
20			55			
21					52	
22		50				
23		42				
24		39				
25					37	
26						35
27						30
28	30					
29					28	
30	24					
31	24					
32					24	
33	22					
34		22				
35		22				
36			21			
37				21		
38		21				
39					20	
40						19
41		19				
42		18				
43		18				
44	17					

Table 12. Document Clustering Taxonomy. (continued)

Clust#	Dynamic			Static		
	Generic	Phys/ Eng/Life	Envir	Generic	Micro-scale	Macro-scale
45		17				
46		17				
47						16
48		16				
49						16
50						16
51					16	
52			15			
53	15					
54						15
55		15				
56						14
57		13				
58					13	
59		12				
60						12
61						12
62						12
63	12					
64					12	
SUM	744	1173	91	336	836	197
TOTSUM		2008			1369	

The 64 clusters cover about 3/4 of the total documents in the database. About 60% can be classified as dynamics, while the remaining 40% can be viewed as statics, subject to the uncertainties due to static-dynamic mixing discussed previously. Dynamics sub-divides into slightly over 1/3 generic (no easily identified associated application), slightly under 2/3 physical/engineering/life sciences, and perhaps 5% environmental sciences. Statics sub-divides into about 25% generic, 60% microscale, and the remaining 15% macroscale.

5. DISCUSSION AND CONCLUSIONS

The initial part of this discussion focuses on the bibliometrics, and the final part focuses on the taxonomies.

The author bibliometrics comparison of 2001–2002 and 1991–1993 showed a substantial regional shift from Europe to Asia over the past decade, and a more moderate shift from universities to research institutes. The regional shift has been noted in other recent text mining studies, and reflects to a large extent the increase in publications output reported by China.

The journal bibliometrics reflected a stronger concentration of Fractals publications in physics journals, with a slight shift in emphasis over the past decade from the more traditional discipline-oriented physics journals to the more generic non-discipline-oriented physics journals. The institutional bibliometrics reflected the shift from European to Asian institutions over the past decade noted under the author bibliometrics, although the shift from universities to research institutes noted under the author bibliometrics was not evident in the institutional bibliometrics results. The country bibliometrics trend over the past decade reflected the regional trend noted above. U.S. co-authorship with China tripled over the past decade, while China's co-authorship with its second largest partner in 1991–1993 (Germany) increased by 50%, and China's co-authorship with its third largest partner in 1991–1993 (Italy) decreased by 80%.

The most cited authors from 2001–2002 have a far different regional distribution from that of the most prolific authors for the same time period. The regional distribution of most cited authors for 2001–2002 resembles more closely the distribution of most prolific authors from 1991–1993. More disconcerting, the list of 18 most prolific authors from 1991–1993 and 20 most highly cited authors had only one name in common. This fact raises the issue of whether an intrinsic incompatibility exists between producing large numbers of papers and producing seminal papers.

The most cited document is a 20-year-old book by Mandelbrot. This is the first time that a book has been the most cited document in the first author's text mining studies. The 10 most highly cited documents were published more than a decade ago! The focus of these documents is on Fractals fundamentals. The highly cited documents in the top 20 list that were published in the mid-1990s reflect the Fractals applications as much as, or more than, intrinsic Fractals fundamentals. These observations suggest a study area whose intrinsic fundamental advances peaked about a decade or two ago, and which has now evolved into an applications focus. This data-based conclusion correlates well with the intuitive conclusion one draws when reading thousands of Fractals Abstracts from the last decade.

Finally, the most cited journal (*Physical Review Letters*) emphasizes rapid publication of 'hot' topics, and would therefore tend to establish primacy in an emerging field. Since one aspect of citations is identifying the original literature of a new topic, a credible journal with these characteristics would tend to receive large numbers of citations. This result should send a clear message to the editors of traditional journals, whose present practices involve long review and publication times, but who wish to improve their Journal Impact Factors.

For taxonomy generation, a combined factor matrix and multi-link word clustering process was developed and used for the first time. The factor matrix served to filter the words input to the clustering algorithm, identified the context-dependent trivial words to be excluded, and identified the

context-dependent words that could be conflated. A document clustering algorithm based on GST text similarity quantification was developed and used for the first time.

In all three clustering approaches, the top-level categorization appeared to consist of a dynamical processes category and a static processes category. At the next categorization level, the factor matrix-word clustering approach showed a sub-division into four categories (Dynamic-Physical/Engineering/Life Sciences, Environmental Sciences; Static-Microscale, Macroscale), while the document clustering approach suggested the addition of a generic category to the static and dynamic second-level categories as well. However, an argument could be made that the two generic categories could be added to the second-level categories from the factor matrix-word clustering approach as well, since the first two clusters from the word clustering dendogram (Clusters A and B) are relatively generic, and could have been placed in such generic categories.

Because of the strong association of Fractals to applications, a few of the lowest level category themes had dynamic and static components. In these few cases, the application played a stronger role in defining the theme for clustering purposes than the dynamic or static characteristic. This was true for medical, environmental, and seismic categories. In these cases, dynamical time series patterns and spatial patterns were important.

Within the focused areas, balance among physical, engineering, and life sciences is reasonable, with less emphasis given to environmental sciences. However, there is a substantial imbalance between the 'hard' and 'soft' sciences. Essentially nothing in the phrase pattern analysis reflected input from the true social and political sciences. The number of articles retrieved from the SSCI was a small fraction of the total articles retrieved. A reading of these 'social science' articles showed minimal effort in the true social and political sciences.

There may be various causes. In the present high-tech world economy, commercial and military, research-sponsoring organizations may be far more interested in pursuing 'hard' science applications of Fractals than 'soft' science applications. Thus, money for social and political science research in Fractals may not be available. Because of potential sensitivities of political and social structure dynamics and trends, Fractals studies may in fact be ongoing, but not published in the open literature. Given the increasing global interest in social group situations and organizations at all levels, including social evolution, political dynamics and evolution, and social network analysis, one would expect that Fractals could provide useful insights for analyzing and predicting social and political trends.

For all practical purposes, applicability of Fractals to the 'softer' sciences remains unexplored.

6. REFERENCES

Cutting, D. R., D. R. Karger, J. O. Pedersen, and J. W. Tukey. 1992. "Scatter/Gather: A Cluster-Based Approach to Browsing Large Document Collections." *Proceedings of the 15th International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR'92), pp. 318–329.

Garfield, E. 1985. "History of Citation Indexes for Chemistry—A Brief Review. *Journal of Chemical Information and Computer Sciences*. vol. 25, no. 3, pp. 170–174.

Guha, S., R. Rastogi, and K. Shim. 1998. "CURE: An Efficient Clustering Algorithm for Large Databases, *Proceedings of the ACM-SIGMOD 1998 International Conference on Management of Data* (SIGMOD'98). 73–84.

Hearst, M. A. 1998. "The Use of Categories and Clusters in Information Access Interfaces." In *Natural Language Information Retrieval*,, T. Strzalkowski, Ed. Kluwer Academic Publishers, Boston, MA.

Karypis, G., E.-H. Han, , and V. Kumar. 1999. "Chameleon: A Hierarchical Clustering Algorithm Using Dynamic Modeling," *IEEE Computer: Special Issue on Data Analysis and Mining*, vol. 32, no. 8, pp. 68–75.

Kostoff, R. N. 1993. "Database Tomography for Technical Intelligence." *Competitive Intelligence Review*, vol. 4, no. 1, pp. 38–43.

Kostoff, R.N. 1994. Database Tomography: origins and applications. *Competitive Intelligence Review. Special Issue on Technology*, vol. 5, no. 1, pp. 48–55.

Kostoff, R. N. et al. 1995. System and method for Database Tomography. U.S. Patent Number 5440481.

Kostoff, R. N., H. J. Eberhart, and D. R. Toothman. 1997a. "Database Tomography for Information Retrieval," *Journal of Information Science*, vol. 23, no. 4, pp. 301–311.

Kostoff, R. N., H. J. Eberhart, D. R. Toothman, and R. Pellenbarg. 1997b. "Database Tomography for Technical Intelligence: Comparative Roadmaps of the Research Impact Assessment Literature and the Journal of the American Chemical Society," *Scientometrics*, vol. 40, no. 1, pp. 103–138.

Kostoff, R. N., H. H. Eberhart, and D. R. Toothman. 1998a. "Database Tomography for Technical Intelligence: A Roadmap of the Near-Earth Space Science and Technology Literature," *Information Processing and Management*, vol. 34, no. 1, pp. 69–85.

Kostoff, R. N. 1998b. "The Use and Misuse of Citation Analysis in Research Evaluation," *Scientometrics* (September), vol. 43, no. 1, pp. 27–43.

Kostoff, R. N., H. J. Eberhart, and D. R. Toothman. 1999a. "Hypersonic and Supersonic Flow Roadmaps Using Bibliometrics and Database Tomography. *Journal of the American Society for Information Science* (April), vol. 50, no. 5, pp. 427–447.

Kostoff, R. N. 1999b. "Science and Technology Innovation," *Technovation*, vol. 19, no. 10, pp. 593–604.

Kostoff, R. N., T. Braun, A. Schubert, D. R. Toothman, and J. A. Humenik. 2000a. "Fullerene Roadmaps Using Bibliometrics and Database Tomography," *Journal of Chemical Information and Computer Science* (January–February), vol. 40, no. 1, pp. 19–39.

Kostoff, R. N., K. A. Green, D. R. Toothman, and J. A. Humenik. 2000b. "Database Tomography Applied to an Aircraft Science and Technology Investment Strategy," *Journal of Aircraft* (July–August), vol. 37, no. 4, pp. 727–730.

Kostoff, R. N. 2000c. "The Underpublishing of Science and Technology results," *The Scientist* (May), vol. 14, no. 9, p. 6–6.

Kostoff, R. N., D. R. Toothman, H. J. Eberhart, and J. A. Humenik. 2001. "Text Mining Using Database Tomography and Bibliometrics: A Review," *Technology Forecasting and Social Change*, vol. 68, no. 3, pp. 223–253.

Kostoff, R. N., R. Tshiteya, K. M. Pfeil, and J. A. Humenik. 2002. "Electrochemical Power Source Roadmaps Using Bibliometrics and Database Tomography," *Journal of Power Sources*, Vol. 110; no. 1, pp. 163–176.

Kostoff, R. N., M. Shlesinger, R. and Tshiteya. 2003. "Nonlinear Dynamics Roadmaps Using Bibliometrics and Database Tomography," *International Journal of Bifurcation and Chaos*. In Press.

MacRoberts, M. H. and B. R. MacRoberts. 1996. "Problems of Citation Analysis," *Scientometrics* (July–August), vol. 36, no. 3,
pp. 435–444.

Prechelt, L., G. Malpohl, and M. Philippsen. 2002. „Finding Plagiarisms Among a set of Programs with Jplag," *Journal of Universal Computer Science*, vol. 8, no. 11, pp. 1016–1038.

Rasmussen, E. 1992. "Clustering Algorithms." In *Information Retrieval Data Structures and Algorithms*, W. B. Frakes and R. Baeza-Yates, Eds. Prentice Hall, Upper Saddle River, N J.

SCI. 2002. Science Citation Index. Institute for Scientific Information. Phila., PA.

Steinbach, M., G. Karypis, and V. Kumar. 2000. "A Comparison of Document Clustering Techniques." Technical Report #00--034. Department of Computer Science and Engineering. University of Minnesota.

Willet, P. 1988. "Recent Trends in Hierarchical Document Clustering: A Critical Review," *Information Processing and Management*, vol. 24, pp. 577–597.

Wise, M. J. 1992. "String Similarity via Greedy String Tiling and Running Karb–Rabin Matching," ftp://ftp.cs.su.oz.au/michaelw/doc/RKR_GST.ps, Dept. of CS, University of Sidney.

Zamir, O. and O. Etzioni. 1998. "Web Document Clustering: A Feasibility Demonstration," *Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR'98). pp. 46–54.

APPENDIX A
GREEDY STRING TILING (GST) CLUSTERING

Greedy String Tiling (GST) clustering is a method of grouping text or text character documents (files) by similarity. All documents to be grouped are placed in a database. Each pair of documents is compared by GST, an algorithm originally used to detect plagiarism (Wise, 1993; Prechelt et al., 2002), and a similarity score is assigned to the pair. Then, hierarchical aggregation clustering (Rasmussen, 1992; Steinbach, Karypis, and Kumar, 2000) is performed on all the documents, using the similarity score for group assignment.

Greedy String Tiling computes the similarity of a pair of documents in two phases. First, all documents to be compared are parsed, and converted into token strings (words or characters). Second, these token strings are compared in pairs for determining the similarity of each pair. During each comparison, the GST algorithm attempts to cover one token string (document) with sub-strings ('tiles') taken from the other string. These sub-strings are not allowed to overlap, resulting in a one to one mapping of tokens. The attribute "greedy" stems from the fact that the algorithm matches the longest sub-strings first.

A number of similarity metrics can be defined once the tiling is completed. One similarity metric is the percentage of both token strings that is covered. Another similarity metric is the absolute number of shared tokens. A third similarity metric is the mutual information index. Depending on the purpose of the matching, additional weightings can be used for the similarity matrix to increase the ranking precision. For example, if plagiarism is one study objective, additional weighting could be given to shared string length. All similarity metrics have positive and negative features, and the choice of metric is somewhat influenced by the study objectives and the structure of the database.

Once the document similarity matrix has been generated, myriad clustering techniques can be used to produce a classification scheme (taxonomy). In the present study, multi-link hierarchical aggregation was used. Three clustering variants were actually generated, although the extension to other clustering schemes is straightforward. Single-link, average-link, and complete-link variants are implemented. The variants differ in how the decision of merging to clusters is made. Single-link requires that the similarity of at least two documents is higher than a certain threshold, while complete-link requires that the similarity between all documents in both clusters be higher than a threshold. Average-link requires that the average pair-wise similarity between the documents of both clusters exceed the threshold. For the present study, average-link appeared to give good results, and was the clustering method used.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-01-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
07-2003	Technical	

4. TITLE AND SUBTITLE

FRACTALS TEXT MINING USING BIBLIOMETRICS AND DATABASE TOMOGRAPHY

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHORS

Dr. Ronald N. Kostoff
Dr. Michael F. Schlesinger
ONR

Guido Malpohl
University of Karlsruhe

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Office of Naval Research
800 North Quincy Street
Arlington, VA 22217–5660

8. PERFORMING ORGANIZATION REPORT NUMBER

SD-ONR 477

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Defense Threat Reduction Agency
8725 John J. Kingman Road MSC 6201
Fort Belvoir, VA 22060-6201

10. SPONSOR/MONITOR'S ACRONYM(S)

ONR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Database Tomography (DT) is a textual database analysis system consisting of two major components: (1) algorithms for extracting multi-word phrase frequencies and phrase proximities (physical closeness of the multi-word technical phrases) from any type of large textual database, to augment (2) interpretive capabilities of the expert human analyst. DT was used to obtain technical intelligence from a Fractals database derived from the Science Citation Index (SCI)/Social Science Citation Index (SSCI). Phrase-frequency analysis by the technical domain experts provided the pervasive technical themes of the Fractals database, and the phrase proximity analysis provided the relationships among the pervasive technical themes. Bibliometric analysis of the Fractals literature supplemented the DT results with author/journal/institution publication and citation data.

15. SUBJECT TERMS

Fractals	text mining	phrase-frequency analysis	information retrieval	citation analysis
bibliometrics	database tomography	phrase-proximity analysis	computational linguistics	clustering

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			R. N. Kostoff
U	U	U	UU	55	19B. TELEPHONE NUMBER *(Include area code)* (703) 696-4198

INITIAL DISTRIBUTION

20012	Patent Counsel	(1)
202753	Archive/Stock	(4)
202752	Library	(2)
2027	M. E. Cathcart	(1)
20275	F. F. Roessler	(1)
202753	D. Richter	(1)

Defense Technical Information Center
Fort Belvoir, VA 22060–6218 (4)

SSC San Diego Liaison Office
C/O PEO-SCS
Arlington, VA 22202-4804

Center for Naval Analyses
Alexandria, VA 22311-1850

Office of Naval Research
ATTN: NARDIC (Code 362)
Arlington, VA 22217–5660

Government-Industry Data Exchange
Program Operations Center
Corona, CA 91718–8000

Army Aeromedical Research Laboratory
Fort Rucker, AL 36362-0577

U.S. Army Research Institute
for the Behavioral and Social Sciences
Alexandria, VA 22333-5600

U.S. Army Research Institute
of Environmental Medicine
Natick, MA 01760-5007

Army Research Laboratory
Aberdeen Proving Ground, MD 21005-5425

Army Research Laboratory
Adelphi, MD 20783 (2)

Aviation and Missile Research, Development,
and Engineering Center
Redstone Arsenal, AL 35898-5000

U.S. Army Communications-Electronics Command
Fort Monmouth, NJ 07703

U.S. Army Communications-Electronics Command
Night Vision Electronic Sensors Directorate
Fort Belvoir, VA 22060-5806

U.S. Army Communications-Electronics Command
Intelligence and Information Warfare Directorate
Fort Monmouth, NJ 07703

Walter Reed Army Institute of Research
Washington, DC 20307-5100

Army Aviation Applied Technology Directorate
Fort Eustis, VA 23604-5577

Air Armament Center
Eglin AFB, FL 32542-6810

Arnold Engineering Development Center
Arnold AFB, TN 37389-9011

Air Force Research Laboratory
Wright-Patterson AFB, OH 45433-7132 (2)

Air Force Office of Scientific Research
Arlington, VA 22203-1954

Defense Advanced Research Projects Agency
Arlington, VA 22203-1714

Armed Forces Radiobiology Research Institute
Bethesda, MD 20889-5603

Naval Air Systems Command
Patuxent River, MD 20670-1547

Naval Medical Research Center
Silver Spring, MD 20910

Naval Research Laboratory
Washington, DC 20375

Naval Undersea Warfare Center
Newport, RI 02841

Naval Air Warfare Center
Weapons Division
China Lake, CA 93555-6100

Naval Surface Warfare Center
Coastal Systems Station
Panama City, FL 32407-7001

Naval Surface Warfare Center
Dahlgren Division
Dahlgren, VA 22448-5100

Naval Postgraduate School
Monterey, CA 93943-5138

Navy David Taylor Center for Maritime
Technology
West Bethesda, MD 20817-5700

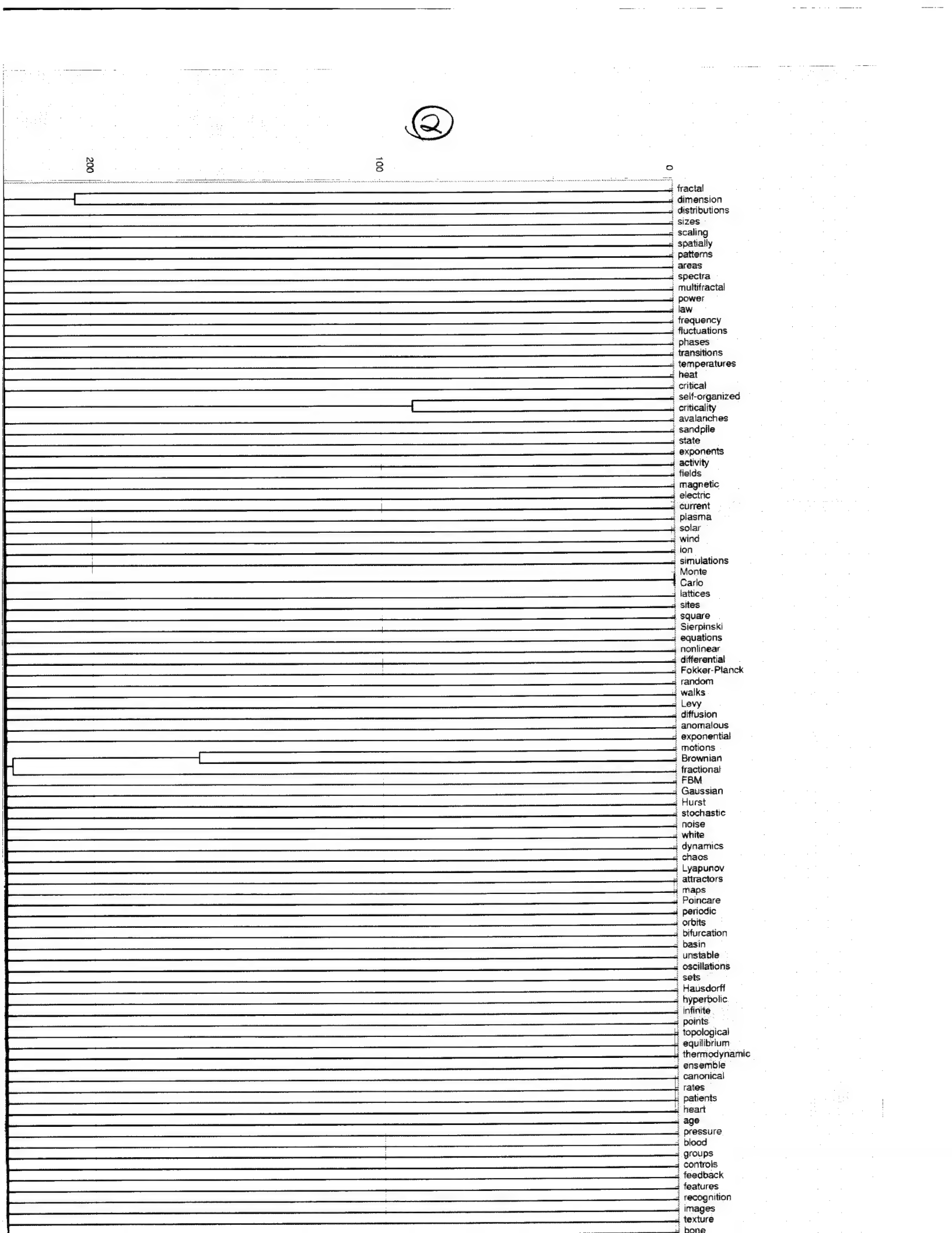
(1)

Distance

500

400

300


200
100
0
fractal
dimension
distributions
sizes
scaling
spatially
patterns
areas
spectra
multifractal
power
law
frequency
fluctuations
phases
transitions
temperatures
heat
critical
self-organized
criticality
avalanches
sandpile
state
exponents
activity
fields
magnetic
electric
current
plasma
solar
wind
ion
simulations
Monte
Carlo
lattices
sites
square
Sierpinski
equations
nonlinear
differential
Fokker-Planck
random
walks
Levy
diffusion
anomalous
exponential
motions
Brownian
fractional
FBM
Gaussian
Hurst
stochastic
noise
white
dynamics
chaos
Lyapunov
attractors
maps
Poincare
periodic
orbits
bifurcation
basin
unstable
oscillations
sets
Hausdorff
hyperbolic
infinite
points
topological
equilibrium
thermodynamic
ensemble
canonical
rates
patients
heart
age
pressure
blood
groups
controls
feedback
features
recognition
images
texture
bone

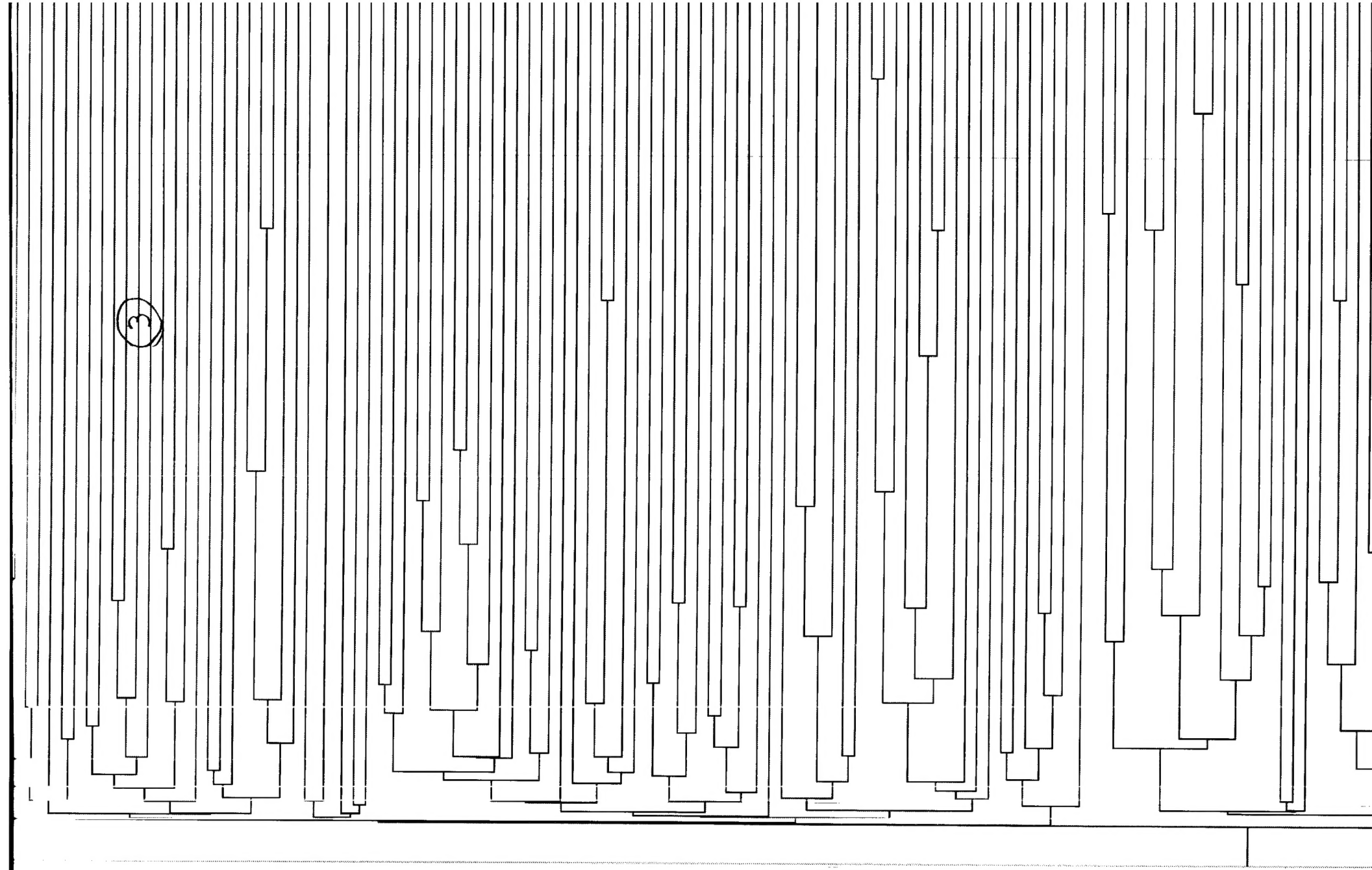
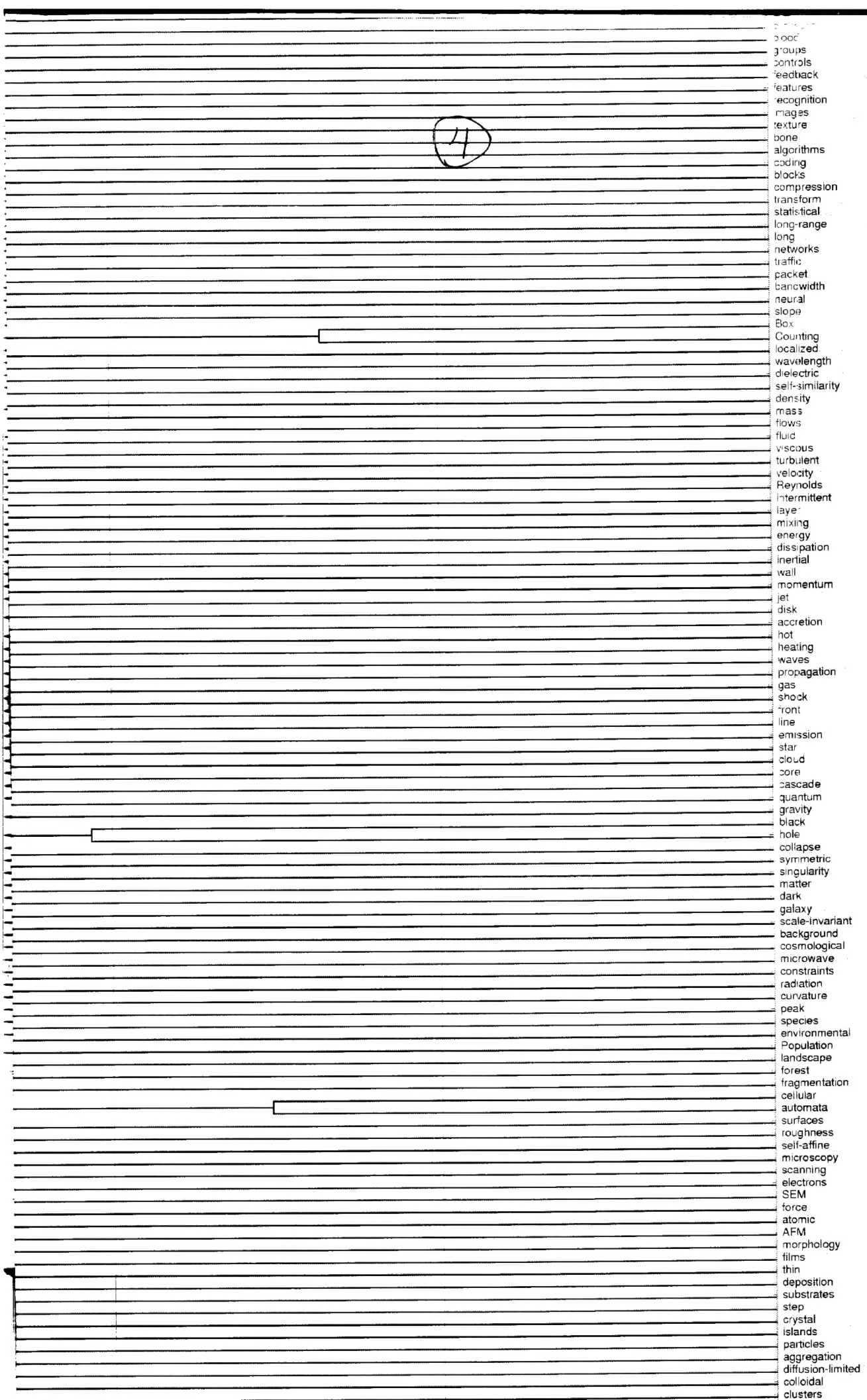
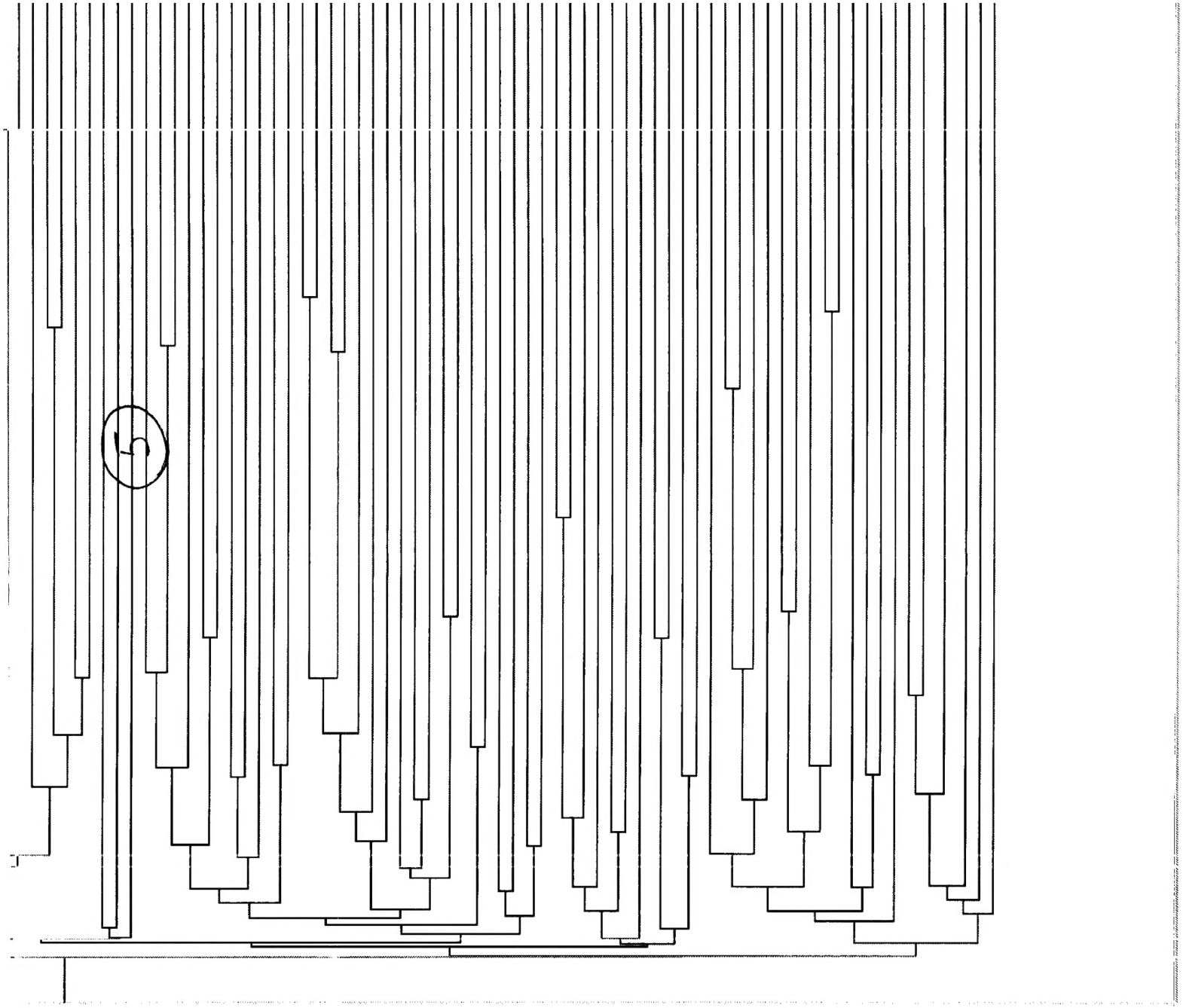
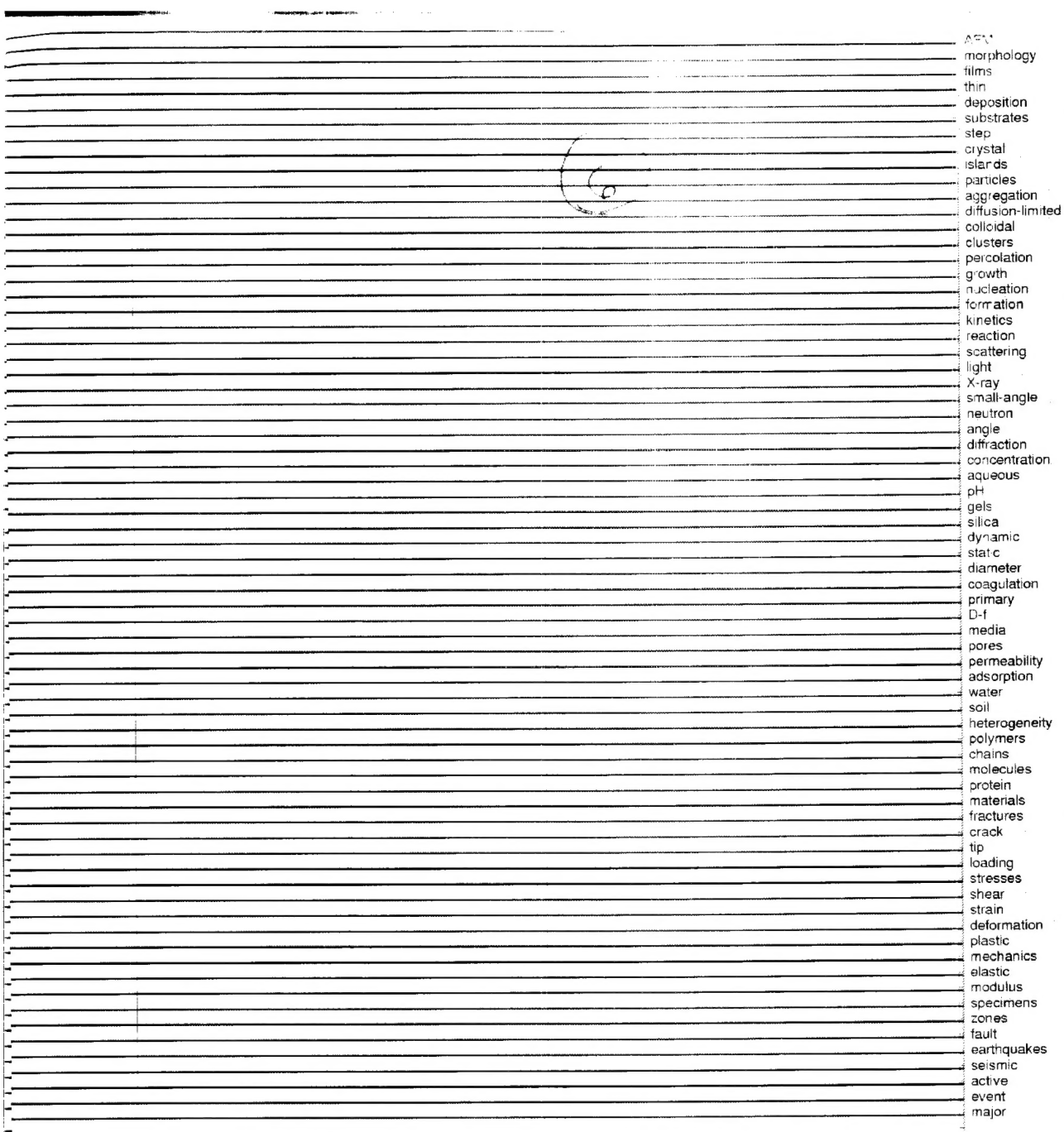


Figure 1. Dendogram


4
blood
groups
controls
feedback
features
recognition
images
texture
bone
algorithms
coding
blocks
compression
transform
statistical
long-range
long
networks
traffic
packet
bandwidth
neural
slope
Box
Counting
localized
wavelength
dielectric
self-similarity
density
mass
flows
fluid
viscous
turbulent
velocity
Reynolds
intermittent
layer
mixing
energy
dissipation
inertial
wall
momentum
jet
disk
accretion
hot
heating
waves
propagation
gas
shock
front
line
emission
star
cloud
core
cascade
quantum
gravity
black
hole
collapse
symmetric
singularity
matter
dark
galaxy
scale-invariant
background
cosmological
microwave
constraints
radiation
curvature
peak
species
environmental
Population
landscape
forest
fragmentation
cellular
automata
surfaces
roughness
self-affine
microscopy
scanning
electrons
SEM
force
atomic
AFM
morphology
films
thin
deposition
substrates
step
crystal
islands
particles
aggregation
diffusion-limited
colloidal
clusters



FO-1/FO-2 blank


AFM
morphology
films
thin
deposition
substrates
step
crystal
islands
particles
aggregation
diffusion-limited
colloidal
clusters
percolation
growth
nucleation
formation
kinetics
reaction
scattering
light
X-ray
small-angle
neutron
angle
diffraction
concentration
aqueous
pH
gels
silica
dynamic
static
diameter
coagulation
primary
D-f
media
pores
permeability
adsorption
water
soil
heterogeneity
polymers
chains
molecules
protein
materials
fractures
crack
tip
loading
stresses
shear
strain
deformation
plastic
mechanics
elastic
modulus
specimens
zones
fault
earthquakes
seismic
active
event
major